PREDICTION IN THERAPEUTIC EFFECTIVENESS RESEARCH: PROLONGED DOSE TITRATION IN WARFARIN PATIENTS AND MODEL TRANSPORTABILITY

Brian Steven Finkelman

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2014

Supervisor of Dissertation

_____

Stephen Kimmel, MD, MSCE
Professor of Medicine and Epidemiology

Graduate Group Chairperson

_____

John Holmes, PhD, Professor of Medical Informatics in Epidemiology

Dissertation Committee

Benjamin French, PhD
Assistant Professor of Biostatistics

Scott Kasner, MD
Professor of Neurology and Emergency Medicine

Stephen Kimmel, MD, MSCE
Professor of Medicine and Epidemiology

Michael Levy, PhD
Assistant Professor of Epidemiology

David Margolis, MD, PhD
Professor of Dermatology and Epidemiology

PREDICTION IN THERAPEUTIC EFFECTIVENESS RESEARCH: PROLONGED DOSE TITRATION IN WARFARIN PATIENTS AND MODEL TRANSPORTABILITY

To my father,

who saved a place on his shelf

for this dissertation

for 28 years.

ACKNOWLEDGMENT

First, I would like to thank my funding sources over the last four and a half years: 5F30HL115992, which was awarded to me by NHLBI; 5R01HL066176, which was awarded to my mentor, Stephen Kimmel, by NHLBI; the NIH MSTP grant for the University of Pennsylvania Perelman School of Medicine, T32-GM07170; and institutional funds from the Center for Therapeutic Effectiveness Research, the Center for Pharmacoepidemiology Research and Training, and the University of Pennsylvania. Completing a dissertation is much easier with a roof overhead and a full stomach.

A huge acknowledgement must be made to my mentor Stephen Kimmel. We began working together the summer following my first year of medical school, and he has always been a wonderful and supportive advisor. His commitment to make time for his students, despite an extremely busy schedule, is really exceptional, and it has made all the difference for me. I would also like to thank my biostatistics advisor, Benjamin French, who encouraged me to always seek the best method, and then made sure that I actually understood it. The rest of my committee—David Margolis, Scott Kasner, and Michael Levy—provided me with invaluable comments on my dissertation, helping me to improve the quality and rigor of the work. A special thank you, also, to Michael Levy, Benjamin French, Ari Friedman, and Elena Prager, who introduced me to data analysis in R, a statistical program that made many parts of this dissertation possible. Thank you also, Elena, for your help with editing this massive document.

Additionally, I would like to acknowledge the staff who helped to ensure that I had the data for a dissertation in the first place: Luanne Bershaw, who probably sees Case Report Forms in her sleep by now; Colleen Brensinger, who turned those forms into a database that was actually

usable; and Sandra Barile, who made sure that all of us showed up to our meetings at the right place and time. I would also like to thank Jennifer Kuklinski and Gabrielle Ostapovich at the Office of Graduate Training in the Center for Clinical Epidemiology and Biostatistics for their help with navigating the maze of graduate degree requirements. Additionally, I would like to thank John Holmes, the director of the PhD program, who has worked tirelessly for many years to help make the PhD program in Epidemiology what it is today. Finally, I want to thank all of the individuals at the MD/PhD program office—and, particularly, Maggie Krall, Maureen Kirsch, and Skip Brass—for providing an impossible amount of logistical and mental support. Without their efforts, I would have probably continued toiling away at my PhD for another year or two before figuring out a way to finish.

Last but not least, I would like to extend thanks to all of the friends and family who supported me over these many years. To my father, Richard, who always made sure that I was working to finish the next paper or grant, and to my mother, Ellen, who always made sure that I did not kill myself in the process. To my sister, Andrea, who always encouraged me to work on research that was meaningful in the real world. And, finally, to my wonderful wife and partner, Elena, for helping me through all the daily struggles that go along with finishing a PhD and for making me look forward to coming home at the end of every long day of work.

ABSTRACT

PREDICTION IN THERAPEUTIC EFFECTIVENESS RESEARCH: PROLONGED DOSE TITRATION IN WARFARIN PATIENTS AND MODEL TRANSPORTABILITY

Brian Steven Finkelman

Stephen Kimmel

Therapeutic effectiveness research relies heavily on prediction modeling, as improving therapeutic outcomes for individuals often requires being able to predict which patients are likely to do poorly on a given therapy. In this dissertation, we examine the specific case of patients starting warfarin therapy, many of whom are at higher risk of bleeding and thrombotic events because they take a long time to determine their optimal therapeutic dose. Additionally, we examine the general problem of transportability of clinical prediction models and whether that problem can be improved through sequential model updating. Specifically, we conducted three projects with the following goals: 1) To determine the social, clinical, and genetic factors associated with time to maintenance dose in patients starting warfarin; 2) To develop and externally validate a prediction model of prolonged dose-titration in these patients; and 3) To determine whether sequential model updating can improve model transportability in a simulation study. Being able to predict which patients are likely to experience prolonged dose titration on warfarin could help clinicians and patients decide whether to use warfarin or a less burdensome alternative oral anticoagulant. Furthermore, the overall utility of this and other clinical prediction models could be greatly increased by strategies that improve model transportability, such as sequential model updating.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

What is therapeutic effectiveness research? The goal of therapeutic effectiveness research is to improve public health by increasing the effectiveness of existing therapies as used in clinical practice. The effectiveness of a therapy is different from its efficacy, which refers to the average effect of a therapy under ideal usage. Efficacy is generally assessed, along with safety, by randomized controlled trials to determine whether therapies should be allowed to be brought to market. Research on therapeutic effectiveness, thus, seeks to identify the factors that lead to the observed discrepancy between a therapy's efficacy and its effectiveness in real-world usage. Therapeutic effectiveness will often depend on a much wider range of factors than efficacy, including clinical factors, such as age, comorbidities, and drug-drug interactions; genetic factors, such as variants in genes related to a drug's pharmacodynamic or pharmacokinetic pathways; and social/behavioral factors, such as access to health care, health literacy, and medication adherence [Bosworth et al., 2011; Ma & Lu, 2011]. As a result, improving the overall effectiveness of a therapy in a population will often necessitate identifying patient subpopulations for whom the therapy is likely to have limited effectiveness, and then utilizing alternative treatment strategies—such as dosing or management changes, interventions designed to improve adherence, or even alternative therapies—in those patients.

The role of prediction modeling in therapeutic effectiveness research. Because improving therapeutic effectiveness often requires identifying patient subpopulations in whom the therapy is generally more or less effective than would be expected in an idealized clinical trial scenario, prediction modeling is of vital importance to therapeutic effectiveness research. Clinical prediction models are most often based on regression methods, in which the outcome of interest—for instance, response to therapy or the development of side effects—is modeled as a

function of several predictor variables, in order to predict the probability of the outcome for a given individual. To be useful clinically, these models must be developed in a rigorous fashion and demonstrate generalizability, or the ability to perform well in the patient population of interest, not just the study cohort used to develop the model. Models are typically assessed both in terms of calibration, which refers to how well predicted probabilities match observed probabilities, and discrimination, which refers to how successful the model is at correctly ranking relatively lower and higher risk individuals. Additionally, model generalizability is typically assessed via external validation, in which the model is tested in a cohort of patients that were not used in the model development process. Finally, it is important to test whether use of the prediction model actually leads to better outcomes in practice. While observational studies can play an important role, testing of prediction model performance is most rigorously done through a randomized controlled trial, comparing outcomes on patients who have been randomized to receive therapy that has been tailored based on the results of prediction models to those who receive standard therapy without prediction. Examples include clinical trials of whether pharmacogenetic dosing algorithms for warfarin led to improvement in anticoagulation control over clinical dosing algorithms or standard clinical practice [Kimmel et al., 2013; Pirmohamed et al., 2013]. Special attention should be paid in such trials to the generalizability of the study population, since effectiveness, not efficacy, is the metric of interest. Furthermore, the time and monetary costs of conducting such trials can often be prohibitive, especially considering that model performance can deteriorate over time, requiring recalibration.

Warfarin is a common oral anticoagulant that has served as a model for therapeutic effectiveness research. Warfarin sodium is a commonly prescribed anticoagulant used for the primary and secondary prevention of thromboembolic disease, and until recently, it was the only available oral anticoagulant in the US [Mohapatra, Tran, Gore, & Spencer, 2005]. The drug has been used in

practice for 60 years; however, it remains difficult to use because of an unusually narrow therapeutic range and as much as a 30-fold variability in dosing requirements for patients to achieve stable therapeutic levels of anticoagulation [Wadelius et al., 2004]. Over-anticoagulation from having too high a dose of warfarin can result in life-threatening bleeding complications, such as intracranial hemorrhage, while under-anticoagulation from having too low a dose of warfarin reduces the efficacy of the therapy, leaving patients at risk for strokes and other thromboembolic events [Higashi et al., 2002; Sconce et al., 2005; White et al., 1987]. Even non-serious adverse events such as minor bleeding can lead to warfarin discontinuation [Gullov, Koefoed, & Petersen, 1999]. As a result of these limitations, much research has been devoted to improving the effectiveness of warfarin therapy in practice. Most of this research has focused on the development of models to predict a patient's required warfarin dose, with the idea that knowing the required therapeutic dose in advance will make it easier to titrate a given patient to a therapeutic level when starting therapy [Gage et al., 2008; Klein et al., 2009]. Pharmacogenetic dosing models for warfarin are typically able to predict within 20% of patients' actual therapeutic dose in about half of individuals [Finkelman, Gage, Johnson, Brensinger, & Kimmel, 2011], although their accuracy has historically been much lower in African Americans [Klein et al., 2009; Limdi et al., 2008; Schelleman, Chen, et al., 2008; Suarez-Kurtz & Botton, 2013].

Current dosing strategies for warfarin often result in a lengthy and dangerous dose titration period. Despite the availability of dosing algorithms, warfarin is still typically dosed empirically, with patients started at the population average dose of 5mg/day and then titrated either up or down based on changes in the international normalized ratio (INR) [Fihn et al., 1993]. As a result, patients often experience a lengthy dose titration period of weeks to months at the onset of warfarin therapy, during which time they are at particularly high risk of complications from improper anticoagulation levels. For instance, it has been estimated that bleeding risk is

approximately 2-6 times higher during the first 3 months of warfarin therapy, and the rate of thromboembolic events has been shown to be elevated very early in a patient's course of warfarin therapy in some contexts, such as following surgery [Brotman, Jaffer, Hurbanek, & Morra, 2004; Fihn et al., 1993]. In addition, a prolonged dose titration phase substantially increases patient burden by increasing the frequency of required visits for INR monitoring for an extended period of time. As a result, such patients may have increased medical costs, reduced quality of life [Dantas, Thompson, Manson, Tracy, & Upshur, 2004], greater dissatisfaction, and higher rates of warfarin discontinuation [Arnsten, Gelfand, & Singer, 1997; Fang et al., 2010], thus depriving these patients of the benefit of a highly efficacious therapy.

Patients at high risk of having a lengthy dose titration period on warfarin therapy may be more appropriately treated with alternative oral anticoagulation agents. In 2010, the FDA approved dabigatran, a direct thrombin inhibitor, for patients with non-valvular atrial fibrillation. Thus, dabigatran became the first oral anticoagulant to be approved in the U.S. since the introduction of warfarin. Rivaroxaban, a Factor Xa inhibitor, was approved by the FDA in November 2011, and another Factor Xa inhibitor, apixaban, was recently approved by the FDA in December 2012. Both dabigatran and rivaroxaban have been shown to be non-inferior to warfarin for prevention of thromboembolic events [Connolly et al., 2009; Patel et al., 2011], while apixaban was shown to be superior to warfarin for stroke prevention in the setting of a randomized trial [Granger et al., 2011]. Bleeding rates were also generally low and either comparable to or lower than warfarin [Siegal & Crowther, 2013]. Moreover, these alternative agents all have the advantage of having much less variability in their dosing requirement for patients [Cove & Hylek, 2013]—although recent evidence suggests that at least dabigatran may have more dosing variability than had been previously thought [Charlton & Redberg, 2014; Cohen, 2014a, 2014b; Moore, Cohen, & Mattison, 2014]—allowing for fixed dosing regimens and eliminating the monitoring burden of

anticoagulation. Furthermore, these newer agents also have fewer food and drug interactions, meaning that they might necessitate fewer lifestyle adjustments and be less prone to fluctuations in anticoagulation levels over the long term.

However, the newer agents have some issues that have prevented them from completely replacing warfarin in clinical practice. All of the drugs are substantially more expensive, as annual direct pill costs for the newer agents are about 60 times more expensive than warfarin [Avorn, 2011]. Furthermore, more of the cost of the newer agents are shifted to patients, since co-pays on the expensive new medications are generally much higher than co-pays for the laboratory testing required with warfarin [Avorn, 2011]. Additionally, dabigatran has shown problems of frequent gastrointestinal side effects and appears to have an increased risk of myocardial infarction relative to warfarin [Ansell, 2010; Uchino & Hernandez, 2012], while rivaroxaban may have an increased risk of spinal hematoma [Jaeger, Jeanneret, & Schaeren, 2011; Steffel & Braunwald, 2011]. Furthermore, it is too soon to know what the full risk profile for apixaban might be in real-world clinical practice.

Ironically, many clinicians have been made uncomfortable by the inability to monitor anticoagulation level in individual patients on the alternative agents. With warfarin, monitoring allowed physicians the opportunity to tailor therapy to those, for example, with increased bleeding risk or renal dysfunction; to identify and potentially address problems with therapy before they led to bleeding or thrombotic events; and to determine whether events that did occur were due to non-therapeutic drug levels [Ansell, 2010]. Removing the frequent contact with health care providers that comes with monitoring might also worsen adherence to the newer anticoagulants [Cutler et al., 2014], and poor adherence could theoretically increase the risk of adverse outcomes for patients on the newer anticoagulants relative to those on warfarin, due to

the shorter half-lives of the newer drugs [Ansell, 2010]. The lack of an antidote to the alternative agents has also led to concern about an inability to stop anticoagulation for patients who develop serious bleeding [Steffel & Braunwald, 2011], including those who are victims of trauma [Cotton, McCarthy, & Holcomb, 2011]; thus, development of antidotes is an active area of current research [Lu et al., 2013]. As a result of all of these issues, there is uncertainty in the clinical community about when to use these newer anticoagulants instead of warfarin [Ansell, 2010; Hankey & Eikelboom, 2010; Kanagasabapathy, Chowdary, & Gatt, 2011; Mangiafico & Mangiafico, 2012].

Our research is motivated by the hypothesis that individual patients who are likely to respond poorly to warfarin could potentially be better treated with less burdensome but more expensive alternative oral anticoagulants, though we will not formally address this specific hypothesis in this dissertation. Recent research has suggested that the cost-effectiveness of dabigatran relative to warfarin is greatest when used in patients who would have had poor INR control on warfarin [Freeman et al., 2011; Shah & Gage, 2011], and there is no reason to expect that this would be different for rivaroxaban and apixaban. Thus, predicting warfarin response in individual patients prior to initiating anticoagulation therapy may be an optimal and cost-effective approach to incorporating alternative oral anticoagulants alongside warfarin in clinical practice.

Existing research is inadequate for identifying patients at high risk of prolonged dose titration on warfarin therapy. While there has been extensive research to determine the factors that affect required therapeutic maintenance dose [Gage et al., 2008, 2004; Kimmel et al., 2008; Klein et al., 2009; Lenzini et al., 2010; Rieder et al., 2005; Schelleman et al., 2010; Schelleman, Chen, et al., 2008; Schelleman, Limdi, & Kimmel, 2008; Voora et al., 2005], much less is known about the factors that lead to a prolonged dose titration phase for patients starting warfarin. Some evidence

suggests that genetic variants associated with maintenance dose may also be associated with prolongation of the dose titration period. For instance, the *APOE* ε3 allele has been associated with delay of reaching maintenance dose in African Americans [Cavallari et al., 2011]. Mutations in *CYP2C9* have also been associated with increased time to maintenance dose [Higashi et al., 2002; Meckley, Wittkowsky, Rieder, Rettie, & Veenstra, 2008], and variants in *VKORC1* have been associated with increased time to first therapeutic INR [Schwarz et al., 2008], although the results for these variants have been mixed [Limdi et al., 2008]. Variants in these genes have also been associated with more frequent dosing changes and greater time spent out of therapeutic INR range [Limdi, Wiener, Goldstein, Acton, & Beasley, 2009; Schwarz et al., 2008]. However, factors that are associated with outcomes in population studies often perform poorly when predicting future outcomes in individuals [Pepe, Janes, Longton, Leisenring, & Newcomb, 2004]. Thus, it is essential to directly test whether these genetic variants could be clinically useful for predicting a prolonged dose titration period in individual patients at the onset of therapy.

Furthermore, given the multifactorial nature of warfarin response, it seems implausible that genetic variants are the only important predictors of a prolonged dose titration phase. However, potentially important clinical and sociodemographic factors have not, to our knowledge, been studied in this context. There is indirect evidence, though, including results from our group, that poor adherence to warfarin could lead to prolongation of the dose titration period, as it has been associated with significantly worse anticoagulation control [Cavallari et al., 2009; Kimmel et al., 2007]. Additionally, we and others have shown that baseline clinical and sociodemographic factors—such as younger age, greater than high school education, current employment, and cognitive impairment—are associated with subsequent poor warfarin adherence [Arnsten et al., 1997; Platt et al., 2008], as has been seen with other medications [Ediger et al., 2007; Kulkarni, Alexander, Lytle, Heiss, & Peterson, 2006; Nikolaus et al., 1996]. However, these prediction

models have not shown very good discrimination in individual warfarin patients [Platt et al., 2010]. Finally, a variety of social and clinical factors have been associated with several other endpoints that may be related to a prolonged dose titration phase, including time in therapeutic INR range, risk of bleeding events, and discontinuation of warfarin therapy [Beyth, Quinn, & Landefeld, 1998; Fang et al., 2010; Gage et al., 2006; Lip, Frison, Halperin, & Lane, 2011; Shireman et al., 2006].

In this dissertation, we aim to improve our ability to predict prolonged dose titration on warfarin therapy as well as better understand its causes. When beginning this research, we hypothesized that baseline clinical, genetic, and social factors could predict prolonged dose titration, which we define as failure to reach stable therapeutic maintenance dose within 3 months of initiating warfarin therapy. In Chapter 2, we focus on identifying both baseline and post-initiation factors that are associated with time to the achievement of maintenance dose. Better knowledge of which factors lead to a longer time to maintenance dose could help clinicians identify patients who are at high risk of prolonged dose titration. Moreover, knowledge of reversible factors that are associated with prolonged dose titration, such as behavioral factors, could potentially even provide targets for interventions designed to improve anticoagulation control in patients on warfarin. In Chapter 3, we focus on developing and externally validating a prediction model for prolonged dose titration when starting warfarin therapy. Accurate prediction of prolonged dose titration could help clinicians decide when to use alternative strategies for anticoagulation, such as less burdensome but more expensive alternative oral anticoagulants, genetic testing to try to improve dosing on warfarin, or more frequent INR monitoring.

Prediction models for individual response to warfarin therapy will need to be able to generalize across a wide variety of clinical settings to maximize their clinical utility. There are over 30

million prescriptions for warfarin in the U.S. every year, with common indications including stroke prophylaxis in atrial fibrillation, the presence of a mechanical heart valve, and treatment for thromboembolic disease [Wysowski, Nourjah, & Swartz, 2007]. Patients on warfarin are managed by specialty anticoagulation clinics, primary care physicians, cardiologists, hematologists, and pharmacists, among others. As a result, it is likely that prediction models developed in one clinical setting may not perform well in other settings, which could diminish their overall usefulness in clinical practice. Deterioration of prediction model performance across different clinical settings is an example of poor model transportability, which is a component of model generalizability that refers to a model's ability to produce accurate and reliable predictions in different populations from the one in which the model was derived [Justice, Covinsky, & Berlin, 1999]. Ultimately, the transportability of a prediction model can only be assessed using validation data from distinct populations.

Utility of clinical prediction models is hampered by concerns about poor transportability across broad areas of clinical medicine. The problem of poor transportability of prediction models is much broader than just predicting warfarin response. For instance, the American Heart Association (AHA) and the American College of Cardiology's (ACC) most recent cholesterol management guidelines were largely dependent on an individual's predicted 10-year risk of cardiovascular events [Stone et al., 2014]. However, the prediction models used in these guidelines have been criticized because of concerns that they over-predict the risk of cardiovascular disease in cohorts other than those used to develop the prediction model [Ridker & Cook, 2013]. Additionally, there are several documented examples of validated prediction models failing to generalize to different populations. For example, the EuroSCORE model, which was developed in European populations to predict 30-day mortality in patients undergoing cardiac surgery, failed to generalize to Australian surgical patients [Yap et al., 2006]. In another example,

a clinical prediction rule for predicting deep vein thrombosis (DVT) performed well in the secondary referral patient population in which it was developed, but failed to generalize to a primary care setting [Oudega, Hoes, & Moons, 2005]. Furthermore, this problem is likely even more widespread because of the many clinical outcomes that are known to vary substantially across clinical sites, including readmission after hospitalization for heart failure [Ross et al., 2008], mortality following surgery for colorectal cancer [Schootman et al., 2014], false-positive results from mammographic screening [Roman, Skaane, & Hofvind, 2014], graft failure after liver transplantation [Asrani et al., 2013], and medication adherence rates among diabetes patients [Sherman, Sekili, Prakash, & Rausch, 2011]. As a result, methods to improve prediction model transportability could be expected to impact a wide range of areas in clinical medicine, and could be especially transformative for therapeutic effectiveness research.

Methods to improve prediction model transportability. Poor transportability of a prediction model often occurs because of a problem of underfitting rather than overfitting [Justice et al., 1999]. In other words, important predictors are either unknown, misspecified, or excluded from the original model, and model performance degrades when tested in new populations with a different conditional prevalence of those predictors. As a result, it can be very difficult to find statistical solutions to problems of transportability using the derivation sample, because by definition, the model needs to be tested on a sample with a different empirical distribution from the derivation sample in order to determine its transportability. Thus, established methods such as Bayesian model averaging [Hoeting, Madigan, Raftery, & Volinsky, 1999], bootstrap aggregation or bagging [Breiman, 1996], and cross-validation [Borra & Di Ciaccio, 2010], which are effective at reducing model overfitting, would not necessarily be expected to lead to improvements in model transportability.

In Chapter 4, we examine sequential model updating of mixed-effects models as a potential strategy for improving prediction model transportability. In this approach, predictions are made on individuals using the best available model at that time. Then, when their outcome data becomes available, the model is re-estimated incorporating the newly available data. In short, sequential model updating solves the problem of derivation datasets not being representative of the population of interest by incorporating data from the population of interest into the derivation dataset over time. In practice, sequential model updating would likely involve integrating the prediction model into an electronic health records system (EHR) that spans multiple clinical sites. Predictions for specific patients could be made using data already available in the EHR, and outcomes would be automatically captured as they occur. This scheme would have the advantage of automatically calibrating to local conditions, thus improving the transportability of the model, without the need to recruit additional cohorts for constructing and validating separate prediction models at each individual site. Our research attempts to quantify these potential gains in prediction accuracy, as well as the types of scenarios where they might be expected to work best. The results of this research could potentially enable future prediction models to be more reliable in real-world clinical practice, both for oral anticoagulation research and for therapeutic effectiveness research in general.

CHAPTER 2. FACTORS AFFECTING TIME TO MAINTENANCE DOSE IN PATIENTS INITIATING WARFARIN

Brian S Finkelman, Benjamin French, Luanne Bershaw, and Stephen E Kimmel

ABSTRACT

Background. Patients starting warfarin often experience lengthy dose-titration periods, when they are at high risk for bleeding and thromboembolism. However, relatively little is known about why some patients take longer than others to reach maintenance dose. Thus, we sought to identify social, clinical, and genetic factors associated with prolonged time to maintenance dose (TTM).

Methods. We conducted a time-to-event analysis, using a prospective cohort of patients initiating warfarin (N = 390). Additionally, we examined whether changes in post-initiation factors were associated with TTM. Finally, we performed a secondary analysis in a subcohort (N = 156) assessing the effect of adherence on TTM.

Results. No genetic or post-initiation factors were significantly associated with TTM. However, previous use of warfarin (HR = 0.64; 95% CI 0.46, 0.88), current smoking status (HR = 0.61; 95% CI 0.39, 0.96), fewer than 4 doctor's visits in the previous year (HR = 0.63 vs 4-12 visits; 95% CI 0.46, 0.88), and worse general health status (HR = 0.63; 95% CI 0.47, 0.84) were significantly associated with longer TTM. Use of illegal injectable drugs (HR = 2.51; 95% CI 1.17, 5.39) was associated with shorter TTM. On secondary analysis, the hazard ratio for better adherence and TTM was 1.70 (95% CI 0.88, 3.27).

Conclusions. Pre-existing behavioral factors, health care utilization, and health quality were associated with TTM in patients initiating warfarin, but clinical comorbidities and genetic factors were not. Future studies are needed to determine whether warfarin patients with prolonged TTM would have better outcomes on alternative agents.

BACKGROUND

Patients initiating warfarin often experience lengthy dose-titration periods of weeks to months, during which time they are at particularly high risk of both bleeding and thromboembolic complications from improper anticoagulation levels [Fihn et al., 1993; Hylek, Skates, Sheehan, & Singer, 1996]. Additionally, during the dose-titration phase, patients may have their international normalized ratio (INR) monitored as frequently as 1-2 times per week, while INR monitoring during the maintenance phase of therapy is generally only once every 1-2 months. As a result of this substantial increase in monitoring burden, patients with a long time to maintenance dose (TTM) may have increased medical costs, reduced quality of life [Dantas et al., 2004], greater dissatisfaction, and higher rates of warfarin discontinuation [Arnsten et al., 1997; Fang et al., 2010]. Furthermore, given the recent availability of alternative oral anticoagulants—including dabigatran, rivaroxiban, and apixaban—a better understanding of the causes of prolonged TTM in warfarin therapy is of increasing importance, because it could potentially help identify patient subsets who might be better treated with alternative agents that, while more costly, do not require monitoring of drug or anticoagulation levels.

In contrast to the large amount of research that has been done on the genetic and clinical factors relating to warfarin maintenance dose requirement [Lee & Klein, 2013], relatively little is understood about the factors that lead to a longer TTM. Previous research on the association between genetic variants and TTM has been mixed [Cavallari et al., 2011; Higashi et al., 2002;

Jorgensen et al., 2009; Limdi et al., 2008; Meckley et al., 2008], with few studies conducted in prospective cohorts. Given the multifactorial nature of warfarin response, however, it seems implausible that genetic variants are the only important factors associated with TTM. Indeed, a variety of non-genetic factors, including social and clinical factors, have been associated with several other endpoints that may be related to prolonged TTM, including poor warfarin adherence [Cavallari et al., 2009; Kimmel et al., 2007], time in therapeutic INR range [Apostolakis, Sullivan, Olshansky, & Lip, 2013; Witt et al., 2009], and risk of bleeding events [Beyth et al., 1998; Gage et al., 2006; Lip et al., 2011; Shireman et al., 2006]. However, such factors have not, to our knowledge, been rigorously studied in the specific context of TTM.

We sought to examine the association between social, clinical, and genetic factors and TTM for patients initiating warfarin. Additionally, we aimed to identify whether changes in factors after warfarin initiation could lead to increased TTM. Identifying such factors could help identify patient subsets that might be better treated with warfarin versus one of the newer anticoagulants. To accomplish these aims, we conducted a time-to-event analysis of the INR Adherence and Genetics (IN-RANGE) cohort, a large prospective cohort of adults initiating warfarin [Kimmel et al., 2007; Platt et al., 2008].

METHODS

IN-RANGE cohort. The IN-RANGE cohort of warfarin patients has been used to study the clinical and genetic predictors of warfarin maintenance dose and adherence [Kealey et al., 2007; Kimmel et al., 2007, 2008; Parker et al., 2007; Platt et al., 2008, 2010; Schelleman et al., 2010, 2007; Schelleman, Chen, et al., 2008]. Participants were recruited from specialty anticoagulation clinics at the Hospital of the University of Pennsylvania (HUP), the Philadelphia Veterans Affairs

Medical Center (PVAMC), and Hershey Medical Center. Institutional review board approval was obtained at all three sites, and all study participants provided written informed consent. Exclusion criteria included being under 21 years old, being unwilling or unable to provide consent, having an abnormal INR prior to starting warfarin or heparin therapy, or the presence of antiphospholipid antibodies. Participants were enrolled between April 2002 and February 2006. All participants in the original IN-RANGE cohort (N = 390) were eligible for inclusion in the current study.

Primary outcome. The primary outcome was the time from warfarin initiation to the first maintenance dose-defining visit, in days. Patients were considered to have achieved maintenance dose if they had three consecutive INRs within the target therapeutic range, with no constraint on the amount of time between INRs. This definition was prespecified prior to cohort enrollment. Having a longer TTM is generally worse for patients because of increases in bleeding and thrombosis risk as well as patient burden. TTM was a secondary outcome of the original IN-RANGE study; however, *a priori* power calculations demonstrated adequate power to detect clinically meaningful hazard ratios (Table 2.1).

Exposures. A total of 38 pre-existing, or 'baseline,' variables were considered for analysis. These included social, clinical, and genetic factors, which were all assessed at the time of recruitment (Tables 2.2 and 2.3). Genetic factors studied were the *VKORC1* -1639G>A variant (rs9923231), the *CYP2C9**2 and *CYP2C9**3 variants (rs1799853 and rs1057910, respectively), and the *APOE* ε2 and ε4 alleles (based on the rs7412 and rs429358 variants, respectively). As described previously [Kimmel et al., 2008], DNA was extracted from buccal swab preparations and analyzed using PCR amplification by collaborators who were blinded to patient characteristics and outcomes. All non-genetic factors were ascertained via self-report, making the data comparable to what would be available to clinicians managing warfarin patients.

Table 2.1. Power calculation for primary analysis

Percent Exposed	Minimum Detectable Hazard Ratio > 1	Maximum Detectable Hazard Ratio < 1
50%	1.4	0.71
35%	1.4	0.71
25%	1.5	0.67
15%	1.6	0.63
10%	1.7	0.59

Calculations are based on a type I error rate of 0.05, 300 subjects reaching maintenance dose, and 80% power. Calculations were performed using PASS 11.

Table 2.2. Baseline social and genetic factors considered as candidate variables for primary analysis and their specifications.

Factor	Specification
Social	
Self-reported race	Binary (0 = not African American; 1 = African American)
Gender	Binary (0 = male; 1 = female)
Marital status	Categorical (1 = married (ref); 2 = separated/divorced; 3 = widowed; 4 = never married)
Employment status	Categorical (1 = working; 2 = unemployed; 3 = retired (ref); 4 = disabled)
Education status	Binary (0 = more than high school; 1 = high school or less)
Annual income per household member	Categorical (1 = < $15,000; 2 = $15,000 to $20,000; 3 = > $20,000 (ref))
Insurance status	Categorical (1 = private (ref); 2 = any VA; 3 = Medicaid; 4 = Medicare only; 5 = no insurance)
Ever used illegal injectable drugs	Binary (0 = no; 1 = yes)
Number of alcoholic drinks per week	Binary (0 = 0–7 drinks; 1 = more than 7 drinks)
Current smoking status	Binary (0 = not current smoker; 1 = current smoker)
Self-reported general health status	Binary (0 = excellent/very good/good; 1 = fair/poor)
No. hospitalizations in past 12 months	Continuous (linear)
No. doctor's visits in past 12 months	Categorical (1 = 0–3 visits; 2 = 4–12 visits (ref); 3 = 13 or more visits)
Had difficulty receiving health care in the past 12 months	Binary (0 = no; 1 = yes)
Genetic	
VKORC1 -1639G>A variant	Binary (0 = no variants; 1 = at least one variant)
*CYP2C9*2* and *CYP2C9*3* variants	Binary (0 = no variants; 1 = at least one variant)
APOE ε2 allele	Binary (0 = no copies; 1 = at least one copy)
APOE ε4 allele	Binary (0 = no copies; 1 = at least one copy)

Table 2.3. Baseline clinical factors considered as candidate variables for primary analysis and their specifications.

Factor	Specification
Clinical	
Age (years) at baseline visit	Continuous (linear)
Body Mass Index	Continuous (linear)
Previous use of warfarin	Binary (0 = no; 1 = yes)
Warfarin indication	Categorical (1 = atrial fibrillation/atrial flutter (ref); 2 = post deep vein thrombosis/pulmonary embolism; 3 = dilated cardiomyopathy/left ventricular thrombosis; 4 = stroke/transient ischemic attack; 5 = other)
Number of interacting medications being used at baseline	Binary (0 = 0–1 medications; 1 = 2 or more medications)
Amiodarone use at baseline	Binary (0 = no; 1 = yes)
Statin use at baseline	Binary (0 = no; 1 = yes)
$CHADS_2$ score	Categorical (1 = 0 (ref); 2 = 1; 3 = 2 or higher)
History of pulmonary embolism	Binary (0 = no; 1 = yes)
History of deep vein thrombosis	Binary (0 = no; 1 = yes)
History of peptic ulcer disease	Binary (0 = no; 1 = yes)
History of gastritis	Binary (0 = no; 1 = yes)
History of stroke	Binary (0 = no; 1 = yes)
History of cancer	Binary (0 = no; 1 = yes)
History of hypertension	Binary (0 = no; 1 = yes)
History of diabetes	Binary (0 = no; 1 = yes)
History of arrhythmia	Binary (0 = no; 1 = yes)
History of congestive heart failure	Binary (0 = no; 1 = yes)
History of myocardial infarction	Binary (0 = no; 1 = yes)
History of any other heart disease	Binary (0 = no; 1 = yes)

Additionally, several 'post-initiation' factors were studied, including changes in the use of interacting medications, quantitative and qualitative changes in diet, changes in weight, and changes in alcohol consumption since starting warfarin. Changes in interacting medications were defined as starting or stopping an interacting medication after warfarin initiation; the list of potentially interacting medications is shown in the Appendix. Finally, warfarin adherence, measured by medication event monitoring system (MEMS) caps [Kimmel et al., 2007], was considered in a secondary analysis because adherence data were only available in 40% of the cohort (N = 156). Some patients did not have MEMS cap data because the devices first became available after enrollment had begun, while others were offered to use the device but declined.

Primary Analysis. Cox regression models, stratified by clinical site, were used for all analyses. Variable selection for the primary model of baseline factors was performed using a combination forward-backward algorithm. Specifically, univariable analyses were performed on baseline candidate variables, and those with $P < 0.2$ via the likelihood ratio test were included in the full model. The variable in the full model with the largest P-value via the likelihood ratio test was successively removed until all P-values were less than 0.1. Next, all previously omitted variables were reintroduced one at a time. Those variables with $P < 0.1$ in this forward step were included in the final model, as were age and race, which were deemed clinically important. The variables included in the final model were age, race, previous use of warfarin, current smoking status, illegal injectable drug use, number of doctor's visits in the previous year, general health status, history of arrhythmia, and having a variant in *VKORC1*. Complete-case analysis was used because only 32 individuals (9% of cohort) were missing data on any of these variables.

To ensure that we could compare the effect of genetic factors with what has previously been observed in the literature, genetic factors were analyzed separately, adjusted for final model variables. Genetic factors were specified as binary variables, indicating whether at least one variant was present, in order to avoid data sparseness when assessing prespecified interactions between genotype and race. For the same reason, *CYP2C9**2 and *3 variants were combined into a single binary variable. The effects of post-initiation factors, adjusted for final model variables, were also analyzed separately. All post-initiation factors were specified as time-dependent variables, with their value representing the total number of changes that an individual had experienced by a given date. Additionally, because of their time-dependent specification, models for post-initiation factors were adjusted for visit number to help prevent confounding by varying frequency of INR monitoring [Fihn et al., 1993].

Finally, because this study used the same cohort for variable selection and model estimation, there was concern about model overfitting and sensitivity to outliers. Thus, all reported point estimates, confidence intervals, and P-values in the primary analysis were estimated using 1,000 bootstrap replications [Efron & Tibshirani, 1994]. Specifically, to perform the bootstrap procedure, individuals were repeatedly sampled with replacement, meaning that the same individuals could be selected multiple times in a given sample. The Cox model was then fit using this bootstrap sample, and hazard ratio estimates were recorded. This procedure was then repeated 1,000 times. Reported hazard ratio point estimates were calculated as the mean hazard ratio estimate from 1,000 bootstrap samples; confidence intervals and P-values were calculated based on the mean and variance of 1,000 bootstrap samples, assuming a normal distribution of the bootstrap samples. This method was chosen to improve the stability and interpretability of stratified estimates based on model interactions; however, use of quantiles from the empirical distribution for producing confidence intervals would have left the results for the main effects essentially unchanged (data not shown). These mean estimates are also slightly more stable than those using model-based estimates in the original sample. Additionally, confidence intervals and P-values are slightly more conservative than what would otherwise be observed.

Secondary Analyses. Warfarin adherence was analyzed using the subcohort of patients with available MEMS cap data (N = 156), adjusting for final model variables. Adherence was specified as a time-dependent binary variable, indicating whether an individual had been ≥80% adherent over the past three visits. Age was excluded from adjusted adherence models to reduce the potential bias from adjustment of near-instruments [Myers et al., 2011; Pearl, 2011], because it is known to be a strong predictor of warfarin adherence [Platt et al., 2008, 2010], while not being associated with the outcome. Use of illegal injectable drugs was also excluded because of unstable estimates due to data sparseness in the subcohort. Finally, we performed a secondary

analysis examining whether individuals with high (≥49 mg/wk) or low (≤21 mg/wk) maintenance dose had increased TTM. As in the primary analysis, point estimates, confidence intervals, and P-values for all secondary analyses were based on 1,000 bootstrap replications.

Sensitivity Analyses. We conducted a sensitivity analysis using inverse probability of censoring weights to determine the potential impact of informative censoring on our results [Cain & Cole, 2009; Robins & Finkelstein, 2000]. In this analysis, a Cox model was constructed with time until censoring, rather than TTM, as the outcome of interest. All candidate baseline variables, post-initiation variables, adherence, visit number, INR, and warfarin dose were considered for inclusion in the model. Factor variables with >1% missingness were given missing indicators, as well, because missing data were felt to be potentially predictive of censoring. Variables were selected using an analogous combination forward-backward algorithm, with less restrictive criteria of $P < 0.25$ for entry and retention. This model was then used to predict individual probabilities of censoring at each patient-visit, which could then be used to construct inverse probability weights, using the formula:

$$w_t = \begin{cases} 0, & C(t) = 1 \\ \dfrac{\Pr(C(t) = 0)}{\Pr\big(C(t) = 0 \big| X(t)\big)}, & C(t) = 0 \end{cases}$$

for which w_t indicates the weight for a patient at time t, $C(t)$ indicates whether an individual was censored at time t, and $X(t)$ indicates an individual's covariates, time-varying or otherwise, at time t. These weights were then applied to the final model in the primary analysis to see how much incorporation of the weights changed the original hazard ratio estimates.

A sensitivity analysis was also performed treating visit number, rather than days, as the unit of time for the primary analysis, in order to look at the impact of potentially variable visit frequencies on our results. Additionally, we performed a sensitivity analysis where standard, non-

bootstrapped model-based estimates were calculated. Finally, the individual effects of *CYP2C9**2 and *CYP2C9**3, as well as using an additive specification (i.e. 0, 1, or 2) for all genetic variants, were assessed in a sensitivity analysis. All analyses were performed using R 3.0.2 [R Development Core Team, 2014].

RESULTS

There were 390 subjects in the cohort, whose characteristics are shown in Table 2.4. Median TTM was 45 days (IQR 15, 135), with 288 subjects (74%) achieving maintenance dose by the end of the study. Median number of visits required to achieve maintenance dose was 7 (IQR 4, 13). Genotype frequencies by race are shown in Table 2.5.

The results for the final model are shown in Table 2.6. Complete data on all variables in the final model were available in 358 subjects (91%), with 267 (75%) achieving maintenance dose by the end of the study. Note that because this is a time-to-event analysis where the "event" is reaching maintenance dose, hazard ratios below 1 indicate that a factor is associated with longer TTM and is worse for patients, on average. This is in contrast to most studies where the event of interest is harmful (i.e. mortality), and hazard ratios below 1 would be considered protective. Previous use of warfarin (HR = 0.64 vs no previous use of warfarin; 95% CI 0.46, 0.88), current smoking status (HR = 0.61 vs current non-smoking status; 95% CI 0.39, 0.96), having fewer than 4 doctor's visits in the previous year (HR = 0.63 vs 4-12 visits; 95% CI 0.46, 0.88), and having fair/poor general health status (HR = 0.63 vs excellent/very good/good general health; 95% CI 0.47, 0.84) were significantly associated with longer TTM. In contrast, use of illegal injectable drugs (HR = 2.51 vs no reported drug use; 95% CI 1.17, 5.39) was associated with shorter TTM.

Table 2.4. Characteristics of the IN-RANGE clinical cohort (N = 390).

Characteristic	N (%) or Mean (SD)	Characteristic	N (%) or Mean (SD)
Age (years)	59.2 (15.0)	*CYP2C9* genotype:	
Female gender	119 (31)	*1*1	283 (76)
Race:		*1*2	59 (16)
African American	174 (45)	*1*3	26 (7)
Caucasian	206 (53)	*2*3	3 (1)
Other	10 (3)	*VKORC1* -1639G>A genotype:	
Body Mass Index:		GG	209 (56)
< 25	122 (32)	GA	149 (40)
25–30	125 (32)	AA	15 (4)
> 30	140 (36)	Insurance status:	
Warfarin indication:		Private	215 (56)
Atrial fibrillation/flutter	188 (48)	Any VA	107 (28)
DVT/PE	116 (30)	Medicaid	16 (4)
DCM/LV thrombosis	26 (7)	Medicare only	17 (4)
Stroke/TIA	22 (6)	None	29 (8)
Other	38 (10)	Employment status:	
Target INR 2–3	389 (99.7)	Working	128 (33)
Maintenance dose (mg/wk)	39.9 (22.0)	Unemployed	34 (9)
Previous use of warfarin	96 (25)	Retired	143 (37)
History of hypertension	192 (49)	Disabled	81 (21)
History of diabetes	107 (27)	Income per household member:	
History of PUD	36 (9)	< $15,000/year	109 (33)
History of CHF	78 (20)	$15,000–$20,000/year	99 (30)
> 1 Interacting medications	210 (54)	> $20,000/year	122 (37)
Smoking status:		AC clinic site:	
Never smoked	141 (36)	HUP	184 (47)
Past smoker	185 (47)	PVAMC	137 (35)
Current smoker	64 (16)	Hershey	69 (18)

Abbreviations:anticoagulation (AC), *congestive heart failure* (CHF), *deep vein thrombosis* (DVT), *dilated cardiomyopathy* (DCM), *Hospital of the University of Pennsylvania* (HUP), *left ventricular* (LV), *peptic ulcer disease* (PUD), *Philadelphia Veterans Administration Medical Center* (PVAMC), *pulmonary embolism* (PE), *and transient ischemic attack* (TIA).

Table 2.5. Frequencies of *VKORC1*, *CYP2C9,* and *APOE* genotypes stratified by race.

Genotype	Not African American N (%)[a]	African American N (%)[a]
VKORC1 -1639G>A		
GG	73 (36)	136 (80)
GA	116 (57)	33 (20)
AA	15 (7.4)	0 (0.0)
CYP2C9		
*1*1	128 (63)	155 (92)
*1*2	47 (23)	12 (7.1)
*1*3	25 (12)	1 (0.6)
*2*3	3 (1.5)	0 (0.0)
APOE		
ε2/ε2	1 (0.5)	4 (2.4)
ε2/ε3	25 (12)	22 (13)
ε2/ε4	3 (1.5)	11 (6.5)
ε3/ε3	131 (64)	80 (47)
ε3/ε4	45 (22)	46 (27)
ε4/ε4	1 (0.5)	7 (4.1)

[a]Percents are rounded to the nearest percent for values ≥10% and to the nearest tenth of a percent for values below that cut-off. As a result, percents may not appear to add up to exactly 100%.

Table 2.6. Unadjusted and adjusted hazard ratios for time to maintenance dose for variables included in the final model.

Baseline Factor[a] (N = 358)[b]	N (%) or Mean (SD)	Unadjusted[c] Hazard Ratio[d]	Unadjusted[c] P-value[e]	Adjusted[c] Hazard Ratio[d]	Adjusted[c] P-value[e]
Age (years)	59 (15)	1.01 (1.00, 1.01)	0.24	1.01 (1.00, 1.02)	0.15
Race					
African American	159 (44)	0.85 (0.65, 1.11)	0.24	1.02 (0.73, 1.42)	0.90
Caucasian or other	199 (56)	—		—	
Previous use of warfarin					
Yes	89 (25)	0.69 (0.52, 0.93)	0.015	0.64 (0.46, 0.88)	0.007
No	269 (75)	—		—	
Current smoking status					
Yes	61 (17)	0.72 (0.47, 1.09)	0.12	0.61 (0.39, 0.96)	0.031
No	297 (83)	—		—	
Self-reported illegal injectable drug use					
Yes	17 (5)	1.65 (0.73, 3.73)	0.23	2.51 (1.17, 5.39)	0.018
No	341 (95)	—		—	
No. doctor's visits in previous year:					
< 4	95 (27)	0.71 (0.52, 0.96)	0.085	0.63 (0.46, 0.88)	0.024
4 – 12	174 (49)	—		—	
> 12	89 (25)	0.86 (0.62, 1.20)		0.88 (0.61, 1.28)	
General health					
Fair/poor	114 (32)	0.66 (0.50, 0.88)	0.005	0.63 (0.47, 0.84)	0.002
Excellent/very good/ good	244 (68)	—		—	
History of arrhythmia					
Yes	189 (53)	0.90 (0.70, 1.16)	0.43	0.79 (0.59, 1.05)	0.10
No	169 (47)	—		—	
No. variants in *VKORC1*					
≥1	159 (44)	1.23 (0.95, 1.59)	0.11	1.33 (0.99, 1.78)	0.061
0	199 (56)	—		—	

[a]All non-genetic factors are based on self-report.
[b]Both unadjusted and adjusted results are from the same complete-case dataset to improve comparability.
[c]All models are stratified by anticoagulation clinic site.
[d]Hazard ratios and confidence intervals are based on the mean and variance from 1,000 bootstrap replications. Hazard ratios less than 1 indicate longer time to maintenance dose; hazard ratios greater than 1 indicate shorter time to maintenance dose.
[e]All P-values are based on the Wald test using the mean and variance of estimates from 1,000 bootstrap replications. Categorical variables were tested jointly.

There was evidence to suggest that the proportional hazards assumption may be violated for our primary analysis ($P = 0.01$), but inspection of survival curves for individual covariates indicated

that this should not have a qualitative effect on our results. The effects of genetic factors alone, stratified by race, are shown in Table 2.7. No genetic variant was significantly associated with TTM either before or after adjustment for covariates (All $P_{main\ effect} > 0.06$), and no significant interactions between genetic variants and race were observed (All $P_{interaction} > 0.4$). As shown in Table 2.8, no post-initiation factor was statistically significant either before or after adjustment for covariates (All $P > 0.2$).

In secondary analyses, better adherence appeared significantly associated with shorter TTM in an unadjusted analysis (HR = 1.95; 95% CI 1.06, 3.59), but it was no longer significant after adjustment for covariates (HR = 1.70; 95% CI 0.88, 3.27), as shown in Table 2.9. By contrast, final maintenance dose was not significantly associated with TTM in either unadjusted [high dose HR = 1.03 (95% CI 0.79, 1.34); low dose HR = 1.13 (95% CI 0.78, 1.64); overall $P = 0.81$] or adjusted [high dose HR = 1.10 (95% CI 0.78, 1.54); low dose HR = 1.11 (95% CI 0.73, 1.69); overall $P = 0.79$] analyses.

In sensitivity analyses, use of inverse probability of censoring weights did not appreciably change the results from those shown in Table 2.6, with a 3.3% mean change in hazard ratio estimates, as shown in Table 2.10. Additionally, use of visit number, rather than days, as the unit of time did not substantially change the results, with a 6.8% mean change in hazard ratio estimates (data not shown). Our results were also not substantially changed when standard, non-bootstrapped estimates were used, with a 1.1% mean change in hazard ratio estimates (data not shown). Finally, use of an additive specification for genetic variants and having separate variables for the *CYP2C9**2 and *CYP2C9**3 variants did not substantially change the results, with small quantitative changes toward the null (data not shown).

Table 2.7. Unadjusted and adjusted hazard ratios for time to maintenance dose for genetic factors, stratified by race.

Genetic Variant (N = 358)[a]	African American	Unadjusted Hazard Ratio[b]	Unadjusted $P_{interaction}$[c]	Adjusted[d] Hazard Ratio[b]	Adjusted[d] $P_{interaction}$[c]
Any *VKORC1*	No	1.09 (0.81, 1.46)	0.42	1.31 (0.93, 1.85)	0.85
	Yes	1.41 (0.78, 2.54)		1.40 (0.71, 2.77)	
Any *CYP2C9*	No	0.97 (0.69, 1.36)	0.99	1.05 (0.73, 1.52)	0.49
	Yes	0.96 (0.53, 1.73)		0.68 (0.35, 1.35)	
Any *APOE* ε2	No	1.08 (0.68, 1.73)	0.93	0.91 (0.52, 1.58)	0.46
	Yes	1.11 (0.62, 2.01)		1.21 (0.61, 2.40)	
Any *APOE* ε4	No	1.01 (0.71, 1.44)	0.93	0.97 (0.67, 1.42)	0.92
	Yes	1.03 (0.57, 1.86)		1.00 (0.51, 1.98)	

[a]Both unadjusted and adjusted results are from the same complete-case dataset to improve comparability.
[b]Hazard ratios and confidence intervals are based on the mean and variance from 1,000 bootstrap replications. Hazard ratios less than 1 indicate longer time to maintenance dose; hazard ratios greater than 1 indicate shorter time to maintenance dose.
[c]P-values for interactions are based on the Wald test using the mean and variance of interaction terms from 1,000 bootstrap replications.
[d]Adjusted for all baseline factors shown in Table 2.6.

Table 2.8. Unadjusted and adjusted hazard ratios for time to maintenance dose for post-initiation factors.

Post-Initiation Factor (N = 358)[a]	Median time to first change[b]	Unadjusted Hazard Ratio[c]	Unadjusted P-value[d]	Adjusted[e] Hazard Ratio[c]	Adjusted[e] P-value[d]
Change in interacting medication	47 (28, 83)	0.93 (0.70, 1.24)	0.62	1.01 (0.76, 1.34)	0.95
Change in diet:					
Qualitative	14 (7, 34)	0.97 (0.80, 1.17)	0.73	1.00 (0.82, 1.23)	>0.99
Quantitative	14 (7, 36)	0.91 (0.78, 1.07)	0.24	0.98 (0.84, 1.15)	0.82
Change in weight	17 (7, 35)	0.93 (0.82, 1.06)	0.26	0.97 (0.83, 1.13)	0.70
Change in alcohol use	50 (29, 86)	0.86 (0.60, 1.23)	0.42	0.96 (0.68, 1.34)	0.80

[a]Both unadjusted and adjusted results are from the same complete-case dataset to improve comparability.
[b]Median time (IQR) in days from the initiation of warfarin to the first change experienced by an individual for the given variable.
[c]Hazard ratios are based on the mean estimate from 1,000 bootstrap replications. Hazard ratios less than 1 indicate longer time to maintenance dose; hazard ratios greater than 1 indicate shorter time to maintenance dose.
[d]All P-values are based on the Wald test using the mean and variance of estimates from 1,000 bootstrap replications. Categorical variables were tested jointly.
[e]Adjusted for all baseline factors shown in Table 2.6, plus visit number to prevent visit frequency from confounding the time-varying covariates.

Table 2.9. Unadjusted and adjusted hazard ratios for time to maintenance dose in subcohort with adherence data.

Factor[a] (N = 143)[b]	Unadjusted		Adjusted (– Adherence)		Adjusted (+ Adherence)[e]	
	Hazard Ratio[c]	P-value[d]	Hazard Ratio[c]	P-value[d]	Hazard Ratio[c]	P-value[d]
≥ 80% adherence[f]	1.95	0.032	—	—	1.70	0.11
	(1.06, 3.59)				(0.88, 3.27)	
African American	0.88	0.60	0.84	0.61	0.90	0.77
	(0.54, 1.43)		(0.44, 1.61)		(0.46, 1.76)	
Previous use of warfarin	0.67	0.12	0.58	0.063	0.59	0.084
	(0.41, 1.11)		(0.32, 1.03)		(0.32, 1.07)	
Current smoker	0.75	0.39	0.68	0.32	0.70	0.37
	(0.39, 1.44)		(0.31, 1.47)		(0.33, 1.52)	
No. doctor's visits in previous year:						
< 4	0.52	0.026	0.47	0.026	0.51	0.053
	(0.32, 0.85)		(0.27, 0.82)		(0.28, 0.91)	
4 – 12	—		—		—	
> 12	0.67		0.68		0.61	
	(0.35, 1.29)		(0.29, 1.57)		(0.27, 1.41)	
Fair/poor general health	0.64	0.055	0.63	0.10	0.69	0.20
	(0.40, 1.01)		(0.36, 1.10)		(0.39, 1.22)	
History of arrhythmia	1.14	0.55	1.01	0.97	1.00	>0.99
	(0.74, 1.78)		(0.57, 1.79)		(0.57, 1.76)	
VKORC1 variant	0.96	0.84	1.06	0.85	1.01	0.97
	(0.62, 1.47)		(0.57, 1.98)		(0.54, 1.88)	

[a]All non-genetic factors, excluding adherence, are based on self-report. Age was excluded from this analysis to prevent over-adjustment, because it is a known strong predictor of warfarin adherence while being very weakly associated with TTM. Illegal injectable drug use was excluded because there were too few self-reported users in the subcohort to produce stable estimates.

[b]Both unadjusted and adjusted results are from the same complete-case dataset to improve comparability; only individuals with adherence data were included in this analysis.

[c]Hazard ratios and confidence intervals are based on the mean and variance from 1,000 bootstrap replications. Hazard ratios less than 1 indicate longer time to maintenance dose; hazard ratios greater than 1 indicate shorter time to maintenance dose.

[d]All P-values are based on the Wald test using the mean and variance of estimates from 1,000 bootstrap replications. Categorical variables were tested jointly.

[e]The adjusted model also included visit number to ensure that visit frequency was not confounding the time-varying covariate.

[f]Adherence was specified in a time-varying fashion, indicating whether the participant had correct adherence on ≥ 80% of the days over the last 3 visits, using medication event monitoring system (MEMS) data.

Table 2.10. Adjusted hazard ratios for time to maintenance dose using inverse probability of censoring weights.

Baseline Factor[a] (N = 358)	Adjusted IPCW Hazard Ratio[b]	P-value[c]
Age (years)	1.01 (1.00, 1.02)	0.16
African American	1.00 (0.72, 1.37)	0.98
Previous use of warfarin	0.66 (0.49, 0.91)	0.011
Current smoker	0.65 (0.44, 0.98)	0.040
Illegal injectable drug use	2.25 (1.20, 4.24)	0.012
No. doctor's visits in previous year:		
< 4	0.68 (0.51, 0.92)	0.038
4–12	—	
> 12	0.94 (0.67, 1.34)	
Fair/poor general health	0.62 (0.46, 0.82)	0.001
VKORC1 variant	1.32 (1.00, 1.74)	0.054

[a]All non-genetic factors are based on self-report.

[b]Hazard ratios less than 1 indicate longer time to maintenance dose; hazard ratios greater than 1 indicate shorter time to maintenance dose. Inverse probability of censoring weights were constructed from a Cox model with covariates including income, difficulty obtaining health care, previous warfarin use, warfarin indication, number of hospitalizations in previous year, number of doctor's visits in previous year, statin use, history of pulmonary embolism, history of congestive heart failure, *VKORC1*, *APOE* $\varepsilon 2$, INR value, visit number, warfarin adherence, and clinic site.

[c]All P-values are based on the Wald test, using robust standard errors to account for the non-independence of the weighted samples. Categorical variables were tested jointly.

DISCUSSION

In this study, we examined the social, clinical, and genetic factors associated with TTM, using the IN-RANGE prospective cohort of adults initiating warfarin therapy. We found that previous use of warfarin, current smoking status, having fewer than 4 doctor's visits in the previous year, and worse general health status were all associated with longer TTM, while use of illegal injectable drugs was associated with shorter TTM. To our knowledge, this study is the first systematic examination of all of these factors for the clinically-relevant outcome of TTM in patients initiating warfarin.

Primary Analysis. Most of the literature on factors associated with TTM has focused on the effects of genetic variants, and our findings for genetic variants are largely consistent with these previous studies. None of the genetic variants studied were significantly associated with TTM. Like other prospective studies [Jorgensen et al., 2009; Limdi et al., 2008], we failed to observe an association between *CYP2C9**2 or *3 and TTM in either African Americans or Caucasians. While evidence suggests that *CYP2C9**5, *6, *8, and *11 may be more important than *CYP2C9**2 and *3 for determining warfarin maintenance dose in African Americans due to their higher prevalence [Cavallari et al., 2010], significant associations between these variants and TTM have not been observed in previous studies [Limdi et al., 2008].

Similarly, *VKORC1* was not significantly associated with TTM in either African Americans or Caucasians, which is consistent with the overall literature [Cavallari et al., 2011; Higashi et al., 2002; Jorgensen et al., 2009; Limdi et al., 2008; Meckley et al., 2008]. Our hazard ratio in African Americans, however, was similar to that observed by Limdi et al. [Limdi et al., 2008], although none of these results were statistically significant. Our study was sufficiently powered to detect clinically meaningful hazard ratios, and even when adjusting for multiple variables we had more than 26 events per degree of freedom in our model, well more than the generally recommended 10 events per degree of freedom [Concato, Peduzzi, Holford, & Feinstein, 1995; Peduzzi, Concato, Feinstein, & Holford, 1995]. Thus, if there is indeed a real effect, it seems likely to be of small magnitude. Finally, our results did not confirm a previous finding of an association between *APOE* and TTM in African Americans [Cavallari et al., 2011]. However, this previous study excluded individuals who did not reach maintenance dose and had limited adjustment for confounders. Therefore, the previous finding could have been the result of selection bias, since many individuals who failed to reach maintenance dose could have had a

prolonged dose titration period, or bias from unmeasured confounding of clinical, social, and behavioral factors.

By contrast, non-genetic factors—including behavioral factors (e.g. smoking status), health care utilization (e.g. number of doctor's visits in the previous year), and health quality (e.g. self-reported general health status)—appeared to be more important than genetic factors for determining TTM (Table 2.6). Worse general health status has been previously shown to be associated with worse warfarin adherence [Platt et al., 2010], and current smoking status has been associated with increased warfarin dose requirement [Gage et al., 2008; Nathisuwan et al., 2011] as well as decreased time in therapeutic range [Apostolakis et al., 2013], so it is unsurprising that these factors were found to be associated with longer TTM. Furthermore, fewer than 4 doctor's visits in the previous year might be a marker for reduced health care access or health literacy, so it could conceivably be related to longer TTM through the effect of these factors on medication adherence and INR monitoring burden. Having fewer doctor's visits in the previous year may also be associated with better general health status; however, the effects of being poorly integrated into the health care system on TTM likely overwhelm any benefits of better health.

More surprising was the finding that previous use of warfarin was associated with longer, rather than shorter, TTM. Previous warfarin users did not differ from new warfarin users in terms of their warfarin indication or comorbidities (data not shown); however, they did appear to have their INRs checked less frequently, with 32% of previous warfarin users being seen at least once per week on average compared to 45% for new warfarin users, although this difference was not statistically significant (Table 2.11). One can hypothesize that physicians may have monitored patients with prior warfarin experience less frequently, thus leading to a longer TTM; however, this explanation likely does not fully explain the observed association, as previous warfarin use

Table 2.11. Association between significant factors and overall visit frequency.

Factor[a] (N = 390)	Median Number of Visits to Maintenance Dose[b]	Overall Visit Frequency[c]		*P*-value[d]
		< 1 per week	≥ 1 per week	
Current smoker:				
No	7 (4, 12)	180 (55)	146 (45)	0.21
Yes	8 (4, 21)	47 (73)	17 (27)	
Illegal injectable drug use:				
No	7 (4, 13)	218 (59)	153 (41)	0.076
Yes	4 (3, 12)	8 (44)	10 (56)	
Previous use of warfarin:				
No	7 (4, 11)	161 (55)	130 (45)	0.12
Yes	8 (4, 16)	65 (68)	31 (32)	
No. doctor's visits in previous year:				
< 4	8 (5, 14)	62 (61)	39 (39)	0.70
4–12	6 (4, 10)	102 (55)	85 (45)	
> 12	7 (3, 24)	61 (62)	37 (38)	
General health:				
Excellent/Very Good/Good	6 (4, 11)	141 (55)	115 (45)	0.39
Fair/poor	9 (4, 16)	81 (64)	45 (36)	0.21

[a]All factors from Table 2.6 that were found to be significantly associated with TTM were included here.
[b]Results are reported as median (IQR).
[c]Results are reported as N (%) for each level of visit frequency for each covariate.
[d]P-values are based on the likelihood ratio test from a logistic regression model, adjusted for anticoagulation clinic site; categorical variables were tested jointly.

was still moderately associated with longer TTM in the sensitivity analysis using visit number, rather than days, as the unit of time.

Similarly, the finding that patients who reported using illegal injectable drugs tended to have a shorter TTM was counterintuitive. While it is possible that physicians were intentionally monitoring these patients more closely, confirmatory evidence will be needed before concluding that the observed association was not primarily due to chance. Changes in post-initiation factors were also not associated with TTM, suggesting either that most of these changes typically do not occur early enough in the course of therapy to have a substantial impact on TTM or that they are

identified by physicians and appropriate dose adjustments are made during the dose titration period. However, changes in post-initiation factors could still be important determinants of anticoagulation control in patients on long-term warfarin therapy after maintenance dose has been achieved and monitoring is typically less frequent. Finally, it is also worth noting that most traditional clinical and demographic factors were not associated with TTM, including all clinical comorbidities examined and use of interacting medications at baseline.

Secondary Analyses. Better adherence was not significantly associated with shorter TTM after adjustment for covariates. However, given that the point estimate for adherence was comparable to significant factors in the primary analysis, it seems plausible that there could be a real effect. Because of their shorter half-lives and inability to monitor, there is some concern that nonadherent patients on alternative oral anticoagulants might be expected to have worse outcomes than nonadherent warfarin patients [Avorn, 2011]. Future studies are needed to clarify the effect of adherence on TTM and the effects of adherence on outcomes with alternative oral anticoagulants.

Limitations. There are several potential limitations of this study: 1) While one strength of our study is that we included all available follow-up time in our analyses, there is still the possibility of bias due to informative censoring. We attempted to assess the impact of informative censoring by performing a sensitivity analysis incorporating inverse probability of censoring weights. Because the results were not appreciably changed, we can be more confident that informative censoring is not substantially biasing our results. 2) Because INRs were checked only at visits to the anticoagulation clinic, there is the potential for interval censoring to bias our results. While the potential bias was small due to visits typically being only about a week or two apart, we attempted to determine the effect of interval censoring through a sensitivity analysis in which

visit number was the unit of time for the analysis. The fact that the results were not substantially changed makes us more confident in the robustness of our results. 3) We were limited to the variables available in this cohort; thus, there may have been other important predictors of TTM that we could not assess or residual confounding of the variables we did examine. For this reason, future studies of TTM will likely need better measurement of social, behavioral, and health care access factors, as well as medication adherence. 4) This study used the same dataset for variable selection and effect estimation, potentially leading to problems with overfitting. To address this issue, we bootstrapped all point estimates and confidence intervals in both primary and secondary analyses. Bootstrapped results were not substantially different from standard estimates; however, these results will still need independent validation. 5) Finally, these data are from specialty anticoagulation clinics, potentially reducing their generalizability to warfarin patients in other clinical settings.

Conclusions. In conclusion, TTM was associated with baseline behavioral factors, health care utilization, and health quality in patients initiating warfarin, while traditional clinical comorbidities and genetic factors appeared less important. The observed associations could plausibly be related to differences in warfarin adherence and visit frequency that occur after warfarin initiation, through their effects on anticoagulation control. Future studies will be needed to address whether warfarin patients with prolonged TTM will have better outcomes on alternative oral anticoagulants.

CHAPTER 3. CAN WE PREDICT PROLONGED DOSE TITRATION IN PATIENTS STARTING WARFARIN?

Brian S Finkelman, Benjamin French, Luanne Bershaw, Colleen M Brensinger,

and Stephen E Kimmel

ABSTRACT

Background. Patients initiating warfarin therapy generally experience a dose-titration period of weeks to months, during which time they are at particularly high risk of both thromboembolic and bleeding events. Accurate prediction of which patients are at higher risk of prolonged dose titration could help clinicians determine which patients might be better treated by alternative anticoagulation therapies that, while more costly, do not require dose titration.

Methods. Prolonged dose titration was defined as having a time to maintenance dose of greater than 12 weeks. The prediction model was derived in a prospective cohort of patients initiating warfarin (N = 390), using a Cox proportional hazards model to account for censoring, and then validated in an external cohort (N = 663). Predictor variables were selected using a modified best subsets algorithm, incorporating cross-validation to reduce overfitting.

Results. Five predictor variables were selected for inclusion in the prediction model: warfarin indication, insurance status, number of doctor's visits in the previous year, current smoking status, and history of congestive heart failure. The AUC of this model in the derivation cohort, as estimated using leave-one-out cross-validation, was 0.66 (95% CI 0.60, 0.74), while in the

external validation cohort, the AUC was only 0.59 (95% CI 0.54, 0.64). Including genetic factors in the model did not improve the AUC (0.59; 95% CI 0.54, 0.64). Examination of relative utility indicated that use of the prediction model was unlikely to provide a clinically meaningful benefit for patients.

Conclusion. Our results suggest that prolonged dose titration cannot be accurately predicted in warfarin patients, at least using traditional clinical, social, and genetic predictors. Our results also highlight the general need for external validation when constructing risk prediction models.

BACKGROUND

Because of the substantial population-level variability in warfarin dose requirement, patients starting warfarin therapy will often experience a lengthy dose-titration period of weeks to months. During this period, they are at particularly high risk of both bleeding and thromboembolic complications from improper anticoagulation levels [Fihn et al., 1993; Hylek et al., 1996]. Patients with a prolonged dose-titration period also face increased burden from more frequent international normalized ratio (INR) monitoring, which can lead to a reduced quality of life and higher rates of discontinuation of a highly efficacious therapy [Arnsten et al., 1997; Dantas et al., 2004; Fang et al., 2010]. Given the availability of less burdensome but more expensive alternative oral anticoagulants [Avorn, 2011]—including dabigatran, rivaroxaban, and apixaban—accurate prediction of which patients are likely to experience a prolonged dose-titration period on warfarin could potentially help clinicians decide when to use warfarin versus one of the alternative agents. Thus, we sought to develop and externally validate a model to predict prolonged dose titration in patients initiating warfarin therapy.

METHODS

Overview. We derived a prediction model for whether a patient initiating warfarin achieved maintenance dose within the first 12 weeks of attempted therapy, using a Cox proportional hazards model. We then validated this model in an external prospective cohort of patients initiating warfarin. All analyses were performed in R 3.1.0 [R Development Core Team, 2014].

Derivation cohort. We derived the prediction model using the IN-RANGE cohort, a large prospective cohort of warfarin initiation that has been used to study the clinical and genetic predictors of warfarin maintenance dose and adherence [Kealey et al., 2007; Kimmel et al., 2007, 2008; Parker et al., 2007; Platt et al., 2008, 2010; Schelleman et al., 2010, 2007; Schelleman, Chen, et al., 2008]. Participants were recruited from specialty anticoagulation clinics at the Hospital of the University of Pennsylvania (HUP), the Philadelphia Veterans Affairs Medical Center (PVAMC), and Hershey Medical Center. Institutional review board approval was obtained at all three sites, and all study participants provided written informed consent. Exclusion criteria were kept to a minimum to ensure patient generalizability. Specific exclusion criteria included being under 21 years old, being unwilling or unable to provide consent, having an abnormal INR prior to starting warfarin or heparin therapy, or the presence of antiphospholipid antibodies. Participants were enrolled between April 2002 and February 2006. All participants in the original IN-RANGE cohort (N = 390) were included as part of the derivation cohort for the current study.

Validation cohort. Once the prediction model was developed, it was then validated in an external cohort. The cohort used for validation was the IN-RANGE2 cohort, which was designed as a follow-up cohort to the original IN-RANGE cohort, with similar data collection methods. Participants were recruited from specialty anticoagulation clinics at HUP, PVAMC, and Johns Hopkins University (JHU). Institutional review board approval was obtained at all three sites, and

all study participants provided written informed consent. Exclusion criteria were purposefully kept similar to the original IN-RANGE cohort, with the only difference being that individuals who were neither Caucasian nor African American (about 3% of the original cohort) were excluded from the IN-RANGE2 study and that the presence of antiphospholipid antibodies was dropped as an exclusion criterion for the IN-RANGE2 study. Participants were enrolled between October 2009 and August 2013. All participants with data available as of August 2014 (N = 663) were included in the validation cohort for the current study.

Primary outcome. The primary outcome was a prolonged dose-titration phase, defined as whether an individual achieved maintenance dose within 12 weeks of attempted warfarin therapy. The 12 week cut-off was selected as a clinically meaningful cut-off, as the first 3 months of warfarin therapy have been shown to be especially high risk for patients [Fihn et al., 1993], and some warfarin indications, such as venous thromboembolism with transient risk factors, often only require a 3 month course of therapy [Agnelli & Becattini, 2008]. Additionally, we used a dichotomous rather than continuous outcome, such as time to maintenance dose, to make it easier for clinicians to incorporate model predictions into their decision-making process. Achievement of maintenance dose was defined as having two consecutive INRs within the therapeutic range, at the same warfarin dose, at least one week apart. Use of this definition allowed for the outcome to be defined the same across both the derivation and validation cohorts. Additionally, the time of maintenance dose achievement was taken as the number of days from warfarin initiation to the first maintenance dose-defining visit in days. Reaching maintenance dose within 4 and 8 weeks were also considered as outcomes in secondary analyses.

Candidate predictors. A total of 28 candidate baseline social and clinical factors were considered for inclusion in the primary prediction model, shown in Table 3.1. Most of these candidate

Table 3.1. Candidate baseline social and clinical predictors and their specifications.

Candidate Predictor	Specification
Social	
Self-reported race	Binary (0 = not African American; 1 = African American)
Gender	Binary (0 = male; 1 = female)
Marital status	Categorical (1 = married (ref); 2 = separated/divorced; 3 = widowed; 4 = never married)
Employment status	Categorical (1 = working; 2 = unemployed/disabled; 3 = retired (ref))
Education status[a]	Binary (0 = more than high school; 1 = high school or less)
Insurance status	Categorical (1 = private (ref); 2 = any VA/Medicare only; 3 = Medicaid/no insurance)
Number of alcoholic drinks per week[a]	Binary (0 = 0 drinks; 1 = 1 or more drinks)
Current smoking status	Binary (0 = not current smoker; 1 = current smoker)
Self-reported general health status	Categorical (1 = excellent/very good (ref); 2 = good; 3 = fair/poor)
No. hospitalizations in past 12 months	Categorical (1 = 0 visits (ref); 2 = 1–2 visits; 3 = 3 or more visits)
No. doctor's visits in past 12 months	Categorical (1 = 0–3 visits; 2 = 4–12 visits (ref); 3 = 13 or more visits)
Had difficulty receiving health care in the past 12 months	Binary (0 = no; 1 = yes)
Clinical	
Age (years) at baseline visit	Continuous (linear)
Body Mass Index[a]	Continuous (linear)
Previous use of warfarin[a]	Binary (0 = no; 1 = yes)
Warfarin indication	Categorical (1 = atrial fibrillation/atrial flutter (ref); 2 = post deep vein thrombosis/pulmonary embolism; 3 = other)
Number of interacting medications being used at baseline[a]	Binary (0 = 0–1 medications; 1 = 2 or more medications)
Amiodarone use at baseline[a]	Binary (0 = no; 1 = yes)
Statin use at baseline	Binary (0 = no; 1 = yes)
$CHADS_2$ score	Categorical (1 = 0 (ref); 2 = 1; 3 = 2 or higher)
History of peptic ulcer disease or gastritis	Binary (0 = no; 1 = yes)
History of stroke	Binary (0 = no; 1 = yes)
History of cancer	Binary (0 = no; 1 = yes)
History of hypertension	Binary (0 = no; 1 = yes)
History of diabetes[a]	Binary (0 = no; 1 = yes)
History of arrhythmia	Binary (0 = no; 1 = yes)
History of congestive heart failure	Binary (0 = no; 1 = yes)
History of myocardial infarction[a]	Binary (0 = no; 1 = yes)

[a]These variables were excluded from the model via a univariable screen, described in the Methods section.

predictors have been previously associated with other warfarin-related outcomes, such as warfarin maintenance dose requirement [Gage et al., 2008; Klein et al., 2009], poor warfarin adherence [Arnsten et al., 1997; Platt et al., 2008, 2010], discontinuation of warfarin [Bushnell et al., 2011; Fang et al., 2010; Song, Sander, Varker, & Amin, 2012], percent time in therapeutic range [Hylek, Heiman, Skates, Sheehan, & Singer, 1998; Kimmel et al., 2007; Wieloch et al., 2011], and risk of bleeding events [Beyth et al., 1998; Gage et al., 2006; Lip et al., 2011; Shireman et al., 2006]. Additionally, after constructing a model from baseline social and clinical factors, we were interested in whether inclusion of genetic factors could improve model prediction. For this analysis, we added genetic variants in *CYP2C9* (rs1799853 and rs1057910) and *VKORC1* (rs9923231), specified in a binary fashion as having at least one variant in the given gene, to the model. These variants were chosen because they have most consistently demonstrated a large association with warfarin maintenance dose in the literature, and are used in the major pharmacogenetic dosing algorithms [Gage et al., 2008; Kimmel et al., 2013; Klein et al., 2009].

Choice of statistical model. Because approximately 11% of the derivation cohort was censored prior to 12 weeks of attempted warfarin therapy, we needed to use a statistical model that could accommodate censoring. As a result, we used a Cox proportional hazards model with time from initiation of warfarin to the achievement of maintenance dose or censoring in days as the outcome. The probability of prolonged dose titration was, thus, the conditional probability of survival predicted from this model at the time-point of interest, 12 weeks of attempted warfarin therapy. Because we were not interested in modeling follow-up time after this cut-off point, all individuals who had not reached maintenance dose by 12 weeks were artificially censored at this time.

Univariable screen. To reduce overall computing time for our analyses to manageable levels, we chose to perform a univariable screen to reduce the number of candidate predictors from the initial 28 to 20, which was determined *a priori* to be an appropriate number of candidate variables, given computational constraints. For each candidate predictor, we constructed a univariable Cox proportional hazards model of the time from initiation of warfarin to the achievement of maintenance dose or censoring. We then estimated the time-dependent area under the ROC curve (AUC) at 12 weeks of follow-up using 10-fold cross-validation for each model. The 20 variables with the best time-dependent AUCs in the univariable screen were selected for inclusion in the modified best subsets variable selection algorithm, described below.

Time-dependent AUC. The time-dependent AUC—developed by Heagerty, et al. [Heagerty, Lumley, & Pepe, 2000]—differs from the standard AUC because it accommodates censoring, and it differs from the commonly used C-index because it assesses model discrimination at a single point in time, rather than over the total duration of follow-up. The time-dependent AUC can thus be interpreted as the probability that a randomly selected individual who has experienced the failure event by time *t* will have a higher predicted probability of failure at time *t* than a randomly selected individual who has not experienced the failure event by time *t*. This statistic is estimated by integrating the time-dependent sensitivity and specificity across all possible cut-off values for the linear predictor derived from the model. Because cross-validation was used during the model development process, the linear predictor was calculated in the data subset that was withheld during estimation of the Cox model, repeated for all data subsets (e.g. 10 times for 10-fold cross-validation). When the model was assessed in the external validation cohort, the linear predictors in that cohort were used without cross-validation.

Because individuals may be censored prior to time *t*, the values for time-dependent sensitivity and specificity need to be estimated from the data. As recommended by Heagerty, et al., we used a nearest neighbor estimator—which is essentially a weighted Kaplan-Meier estimator based on a nearest neighbor kernel function, developed by Akritas [Akritas, 1994]—which allows for monotonicity of sensitivity and specificity and for the censoring process to depend on the predictive marker of interest. This estimator is dependent on a smoothing parameter, λ, where 2λ represents the percentage of observations that are included in an individual observation's neighborhood; in our case, we chose the default value of $\lambda = 0.025$. The "survivalROC" package in R was used to facilitate these calculations [Heagerty & Saha-Chaudhuri, 2013].

Variable selection algorithm. Variable selection was conducted using a modified best subsets algorithm [Miller, 2002]. This algorithm was designed to optimize model discrimination, or how well a model distinguishes between those who did and did not experience the outcome (in this case, those who had a prolonged vs non-prolonged dose-titration phase, respectively). We calculated the time-dependent AUC at 12 weeks using 10-fold cross-validation for all possible combinations of the candidate predictors up to 10 predictor variables in length (616,665 combinations) to reduce our chances of selecting a combination based on overfitting. Because we felt that leave-one-out cross-validation (LOOCV)—in which one person at a time is removed from the dataset to build the model and then used for model testing, for all individuals in the dataset—was a better estimate of external validation than 10-fold cross-validation [Hastie, Tibshirani, & Friedman, 2009], we opted to estimate the time-dependent AUC using LOOCV in the 1,000 best models based on 10-fold cross-validation for each subset size (8,210 combinations). The combination of predictors that led to the highest time-dependent AUC using LOOCV was then selected as our final prediction model.

In short, our algorithm was designed to select the combination of candidate variables with the best estimated LOOCV time-dependent AUC. Furthermore, this strategy had the advantage of choosing the best subset based on LOOCV, without the nearly 40-fold increase in computing time that would be required by calculating the time-dependent AUC using LOOCV in all possible combinations of predictors. A sensitivity analysis showed that this algorithm selected the exact same best combination of predictor variables as using LOOCV on all possible combinations up to 6 predictor variables in length. Once selected, prediction model variables were then inspected graphically to ensure proper functional form, and all coefficients were examined to ensure that the direction of effect reported by the model was consistent with the available literature.

Linear shrinkage factor. Because regression coefficients are often overestimated in small samples, prediction models will often show better calibration for out-of-sample predictions when coefficients are shrunk toward zero [Van Houwelingen & Le Cessie, 1990]. Thus, we sought to apply a linear shrinkage factor—which has been shown to perform well in small samples for improving model calibration, without sacrificing model discrimination [Steyerberg, Eijkemans, Harrell, & Habbema, 2000]—to our final prediction model. To estimate the shrinkage factor, we fit the model in a bootstrap sample of the derivation cohort. We then calculated the linear predictors of the individuals in the derivation cohort using the model coefficients from the bootstrap sample. The slope of the actual observed outcomes regressed on these bootstrapped linear predictors could then be used as an estimate of the shrinkage factor. To form a stable estimate of the shrinkage factor, we calculated the mean slope over 1,000 bootstrap replications. All of the original model coefficients were then multiplied by this shrinkage factor to produce the final shrunk coefficients, which were used for generating predictions in the external validation cohort. Because all of the coefficients are being multiplied by the same factor, the rank order of individual predictions is preserved and model discrimination is not affected by shrinkage.

In order to ensure that shrinkage was toward the overall mean and not toward the overall reference category, continuous variables needed to be centered at the mean and categorical variables had to be coded using simple contrasts. In this contrast method, reference groups were coded as $-1/k$, while non-reference categories were coded as $(k-1)/k$, where k is the number of categories. In this contrast method, the reference category of 0 is equivalent to the overall mean of the sample in which the model is being fit. Note that the difference between the reference and non-reference categories is still 1; thus, the interpretation of coefficients in this contrast method is identical to the more common dummy coding for categorical variables (i.e. 0 for reference and 1 for non-reference categories).

Model assessment and validation. The final prediction model was then assessed in a separate validation cohort, described above. Predictions from the model were used to estimate the time-dependent AUC as the primary measure of model discrimination in the validation dataset. Additionally, genetic predictors were added to the model to see if there was a significant difference in the AUC between the two models. The integrated discrimination improvement (IDI) between the two models was also estimated [Liu, Kapadia, & Etzel, 2010]. We also assessed the calibration of the prediction model using calibration plots. Finally, we examined the clinical utility of the prediction model using decision curves and plots of the relative utility of the model versus the risk threshold [Baker, Cook, & Vickers, 2009; Baker, 2009; Vickers, Cronin, Elkin, & Gonen, 2008; Vickers & Elkin, 2006]. Confidence intervals for all estimates were generated using the 2.5^{th} and 97.5^{th} percentiles of estimates in 1,000 bootstrap replications.

The methods for determining clinical utility rely on the concept of the risk threshold, which is the probability of the outcome at which the clinician is indifferent about which treatment strategy to use; in other words, it is the probability at which the costs of false positive and false negative

mistakes are equal [Pauker & Kassirer, 1975]. Furthermore, the consequences of basing a clinical decision on the predicted probability from a risk prediction model can be estimated as a function of the risk threshold. While the exact threshold will vary depending on the value that physicians and patients place on certain outcomes, the metric can be used to determine the clinical usefulness of a given model under a range of possible thresholds. For our prediction model, given broadly similar safety and efficacy profiles for warfarin and the alternative anticoagulants (with the possible exception of apixaban) [O'Dell, Igawa, & Hsin, 2012; Rollins, Silva, Donovan, & Kanaan, 2014], the risk threshold for a given patient would likely depend primarily on his or her relative costs of INR monitoring on warfarin versus the out-of-pocket financial costs of the alternative anticoagulant agents. In this scheme, patients that are more burdened by financial costs would have a risk threshold above 0.5, while those that are more burdened by INR monitoring would have a risk threshold below 0.5.

Decision curves plot the net benefit of various treatment strategies versus the risk threshold, where the net benefit is equal to the true positive rate minus the false positive rate, weighted as a function of the risk threshold [Vickers & Elkin, 2006]. In this case, the net benefit is calculated relative to the strategy of using standard warfarin therapy in all patients. The curve shows the values of the risk threshold where using the prediction model would be expected to provide a net benefit above the strategies of using the same treatment in every patient. Relative utility is a related measure of the usefulness of a prediction model that is essentially a rescaling of the net benefit, and it can be interpreted as the net benefit of the prediction model, compared to using the same treatment strategy in all patients, as a fraction of the net benefit of perfect prediction [Baker et al., 2009]. A relative utility of 1 indicates that the model performs as well as perfect prediction, while negative values indicate that the model is worse than using the same strategy in everyone.

RESULTS

The characteristics of the derivation and validation cohorts are shown in Table 3.2. The overall prevalence of prolonged dose-titration was 30% in the derivation cohort and 38% in the validation cohort. The variable selection algorithm found that the best LOOCV time-dependent AUC was in a model with the following five variables: warfarin indication, insurance status, number of doctor's visits in the previous year, current smoking status, and history of heart failure. The LOOCV time-dependent AUC in this model was estimated as 0.66 (95% CI 0.60, 0.74). A comparison of this model to the other top performing models with different numbers of predictor variables, as measured with cross-validation, suggested that using cross-validation successfully avoided complex models that were more accurate merely because of having extra degrees of freedom (Figure 3.1). The shrinkage factor based on 1,000 bootstrap replications was estimated to be about 0.82, indicating a moderate degree of overfitting in the original model. Coefficients from the final prediction model, after applying the linear shrinkage factor, are shown in Table 3.3.

When tested in the validation cohort, the AUC of the prediction model at 12 weeks was 0.59 (95% CI 0.54, 0.64). The ROC curve for this model is shown in Figure 3.2. The AUC of the model at 8 weeks was 0.57 (95% CI 0.53, 0.62) and at 4 weeks was 0.57 (95% CI 0.52, 0.62). The calibration of the main model was examined by comparing predicted probabilities to observed frequencies across risk deciles (Figure 3.3); the Hosmer-Lemeshow test for goodness of fit did not show significantly poor model calibration ($P = 0.73$). Addition of genetic factors did not significantly change the AUC at 12 weeks ($P > 0.99$), with the point estimate remaining unchanged at 0.59 (95% CI 0.54, 0.64). A comparison of the ROC curves for the models with and without genetic factors is shown in Figure 3.4. The calibration of the genetic model, however, seemed worse than the main model, though the level of miscalibration was not significantly worse than what would be expected due to chance, using the Hosmer- Lemeshow test ($P = 0.06$).

Table 3.2. Characteristics of the derivation and validation cohorts.

Variable	Derivation cohort (N = 390)[a]	Validation cohort (N = 663)[a]	P-value[b]
Age			
< 45	65 (17)	135 (20)	< 0.001
45 – 55	74 (19)	131 (20)	
55 – 65	103 (26)	219 (33)	
65 – 75	83 (21)	116 (18)	
75+	65 (17)	60 (9)	
Female gender	119 (31)	250 (38)	0.02
African American race	174 (45)	466 (71)	< 0.001
Body Mass Index			
< 25	122 (32)	186 (28)	0.11
25 – 30	125 (32)	189 (29)	
> 30	140 (36)	280 (43)	
History of hypertension	192 (49)	461 (70)	< 0.001
History of diabetes	107 (27)	190 (29)	0.71
History of peptic ulcer disease	36 (9)	98 (15)	0.01
History of heart failure	78 (20)	141 (21)	0.65
Warfarin indication			
AFib/AFlutter	188 (48)	214 (32)	< 0.001
DVT/PE	116 (30)	343 (52)	
Other	86 (22)	105 (16)	
Previously used warfarin	96 (25)	209 (32)	0.02
Smoking status			
Never	141 (36)	275 (42)	< 0.001
Past	185 (47)	235 (36)	
Current	64 (16)	148 (22)	
Insurance status			
Private	215 (56)	276 (42)	< 0.001
VA/Medicare/Other	124 (32)	272 (41)	
Medicaid/None	45 (12)	110 (17)	
Employment status:			
Working	128 (33)	167 (25)	< 0.001
Unemployed	34 (9)	49 (7)	
Retired	143 (37)	192 (29)	
Disabled	81 (21)	251 (38)	
Annual income:			
< $15,000	109 (33)	228 (41)	< 0.001
$15,000 - $20,000	99 (30)	45 (8)	
> $20,000	122 (37)	282 (51)	
Site:			
HUP	184 (47)	263 (40)	< 0.001
PVAMC	137 (35)	198 (30)	
Hershey	69 (18)	—	
JHU	—	202 (30)	

[a]All values are reported as N (%).
[b]P-values are based on the chi-square test.

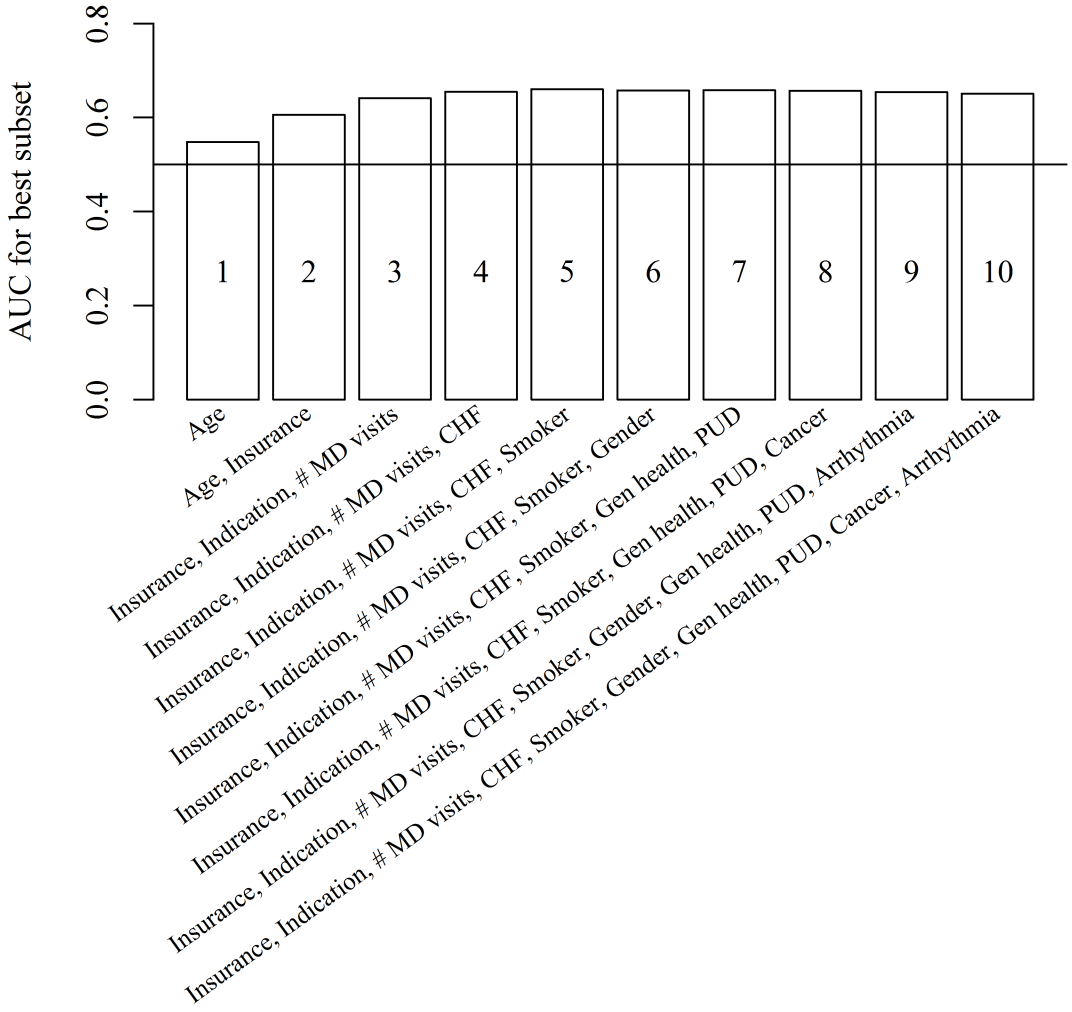

Figure 3.1. Comparison of best prediction models by number of predictor variables in the model. *Prediction models compared by the time-dependent AUC at 12 weeks, as estimated by leave-one-out cross-validation (LOOCV).*

Table 3.3. Final prediction model coefficients.

Predictor variable	Shrunk coefficient[a,b]
Warfarin indication	
AFib/Aflutter	—
DVT/PE	-0.47
Other	-0.33
Insurance status	
Private insurance	—
VA/Medicare	-0.14
Medicaid/None	-0.42
Number MD visits in previous year	
<4	-0.29
4-12	—
>12	-0.23
Current smoker	-0.17
History of heart failure	-0.21

[a]Coefficients were multiplied by a linear shrinkage factor, equal to about 0.82, based on 1,000 bootstrap replications.
[b]Negative coefficients indicate a higher probability of prolonged dose titration.

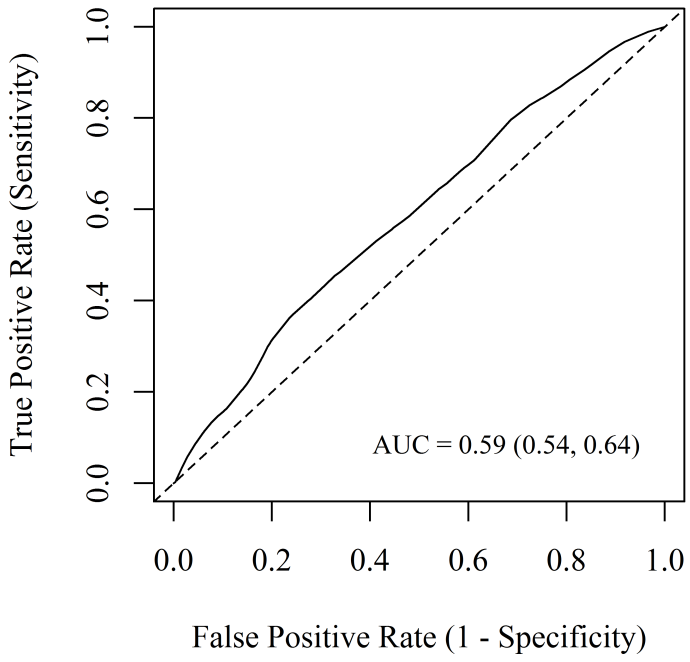

Figure 3.2. ROC curve for the prediction model as tested in the validation dataset.

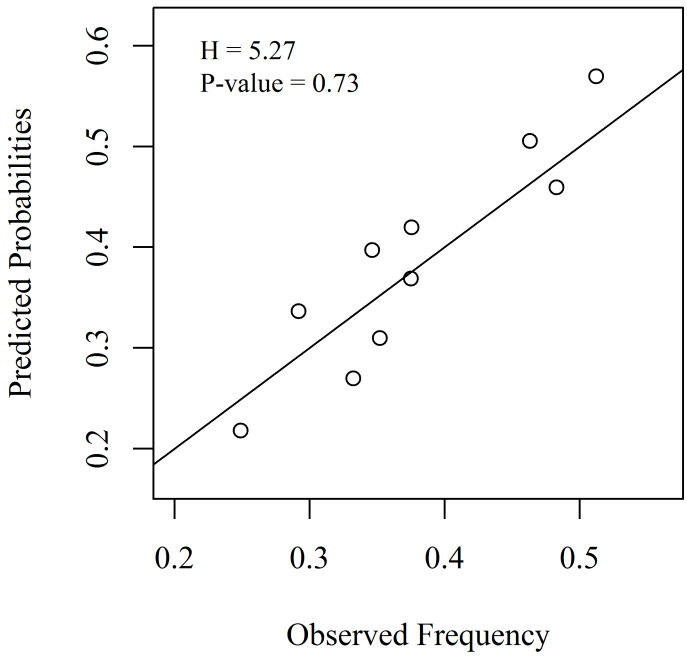

Figure 3.3. Predicted probability vs observed frequency of prolonged dose titration by risk decile.

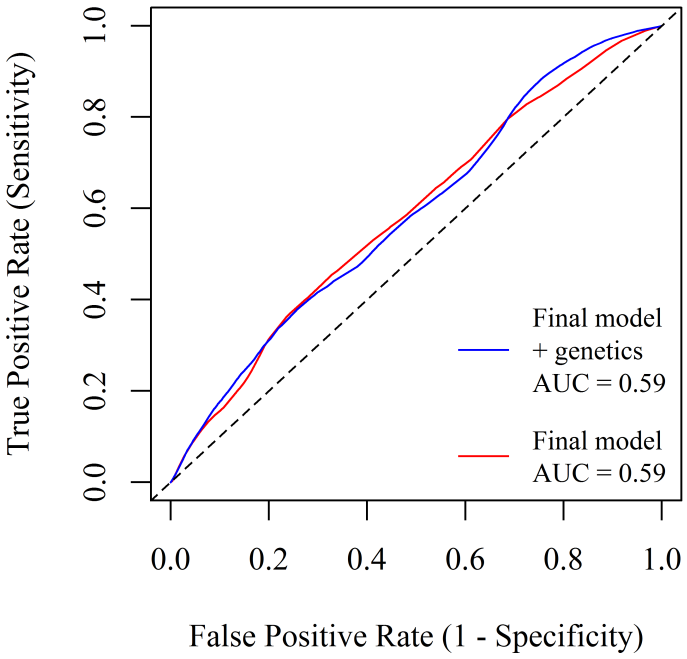

Figure 3.4. Comparison of ROC curves for the prediction models with and without the addition of genetic factors.

The integrated discrimination improvement from adding genetic factors to the model was estimated as 0.01 (0.00, 0.02), which is equivalent to a 7% increase in model discrimination over the model without genetic factors.

To examine the clinical utility of the prediction model, the sensitivity, specificity, and positive and negative predictive values for various risk thresholds were calculated (Table 3.4). Similarly, Figure 3.5 shows the relationship between the positive and negative predictive values and the proportion that are classified as positive across the full range of risk thresholds. Predicted probabilities of prolonged dose titration in the validation cohort ranged from about 16% to 63%; thus, predictive values only varied over this range. The relative utility of the model—which can be understood as the net benefit of the current model, compared to not using a prediction model, as a fraction of the net benefit of perfect prediction—across the full range of risk thresholds is shown in Figure 3.6. The maximum relative utility observed was 9.4%, and the relative utility was negative for the risk threshold range of 48% to 62%. Comparisons of relative utility and decision curves for the models with and without genetic factors are shown in Figures 3.7 and 3.8, respectively.

We also examined site-specific differences in model performance in *post-hoc* analyses. Differences in the characteristics of the derivation and validation cohorts at HUP and PVAMC are shown in Table 3.5. The time-dependent AUC at 12 weeks was 0.60 (95% CI 0.51, 0.67) at HUP, 0.55 (95% CI 0.45, 0.63) at PVAMC, and 0.61 (95% CI 0.53, 0.69) at JHU. Finally, the observed frequency of prolonged dose titration was 32%, 34%, and 48% at HUP, PVAMC, and JHU, respectively; predicted probabilities of the outcome at the respective sites, however, were 37%, 39%, and 38%.

Table 3.4. Model characteristics at various risk thresholds.

Risk threshold[a]	C_{FP}/C_{FN}[b]	Sensitivity	Specificity	Positive Predictive Value	Negative Predictive Value	Proportion Predicted Positive
10%	0.11	1.00	0.00	0.37	—	1.00
20%	0.25	0.99	0.03	0.38	0.87	0.98
30%	0.43	0.83	0.28	0.41	0.74	0.76
40%	0.67	0.52	0.58	0.43	0.67	0.46
50%	1	0.19	0.86	0.44	0.64	0.16
60%	1.5	0.03	0.99	0.57	0.63	0.02
70%	2.33	0.00	1.00	—	0.63	0.00

[a]The risk threshold refers to the cut-off probability, where one classifies individuals as positive when predicted to be above the cut-off or negative when predicted to be below the cut-off. In this case, being "positive" refers to having a high probability of prolonged dose titration on warfarin, potentially leading a physician to choose an alternative therapy.
[b]C_{FP}/C_{FN} refers to the ratio of the costs of false positive and false negative mistakes that are implied by the risk threshold, according to decision theory.

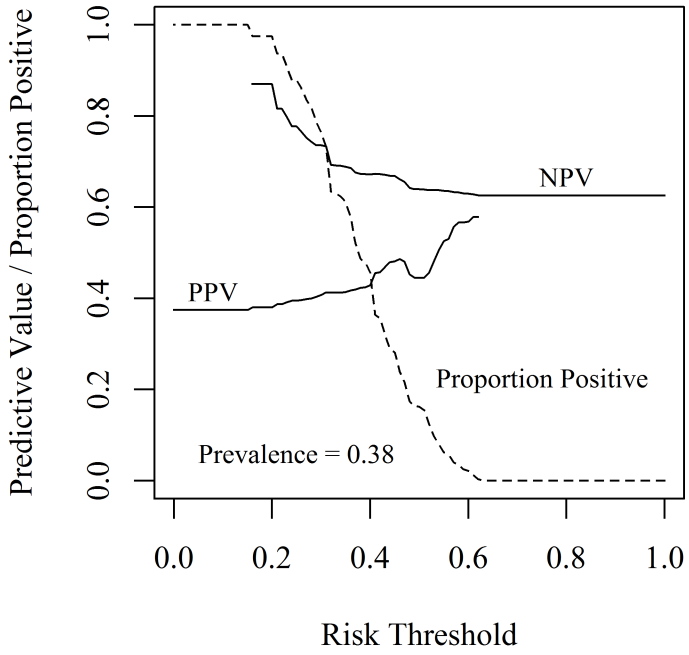

Figure 3.5. Positive predictive value, negative predictive value, and proportion of patients classified as positive across the range of values for the risk threshold. *Individuals with a predicted probability of prolonged dose titration are classified as positive. The absence of a curve in a given region indicates that the measure is undefined in that region; for instance, positive predictive value is undefined when no patients are classified as positive.*

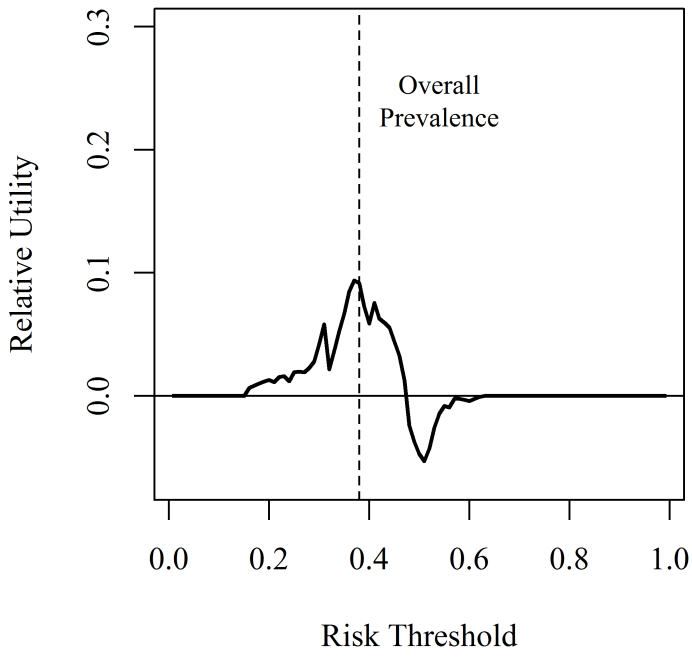

Figure 3.6. Relative utility of the prediction model across the full range of risk thresholds.

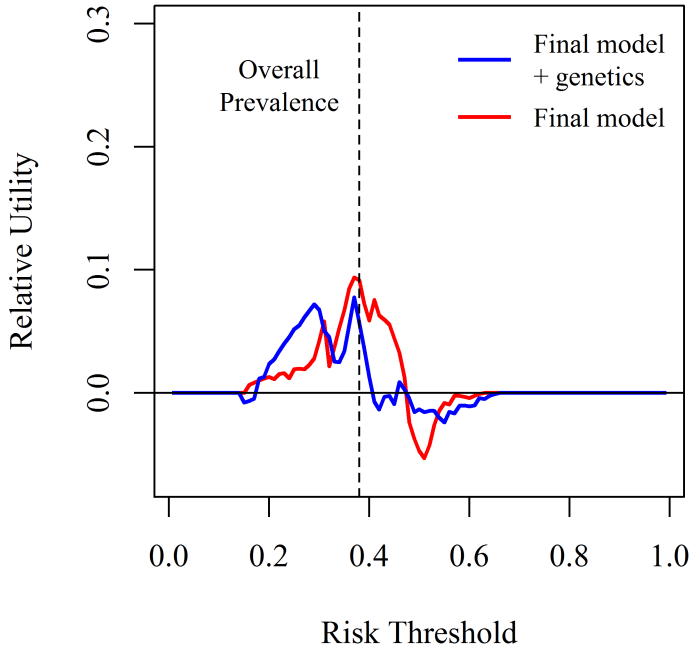

Figure 3.7. Comparison of relative utility curves in prediction models with and without genetic factors.

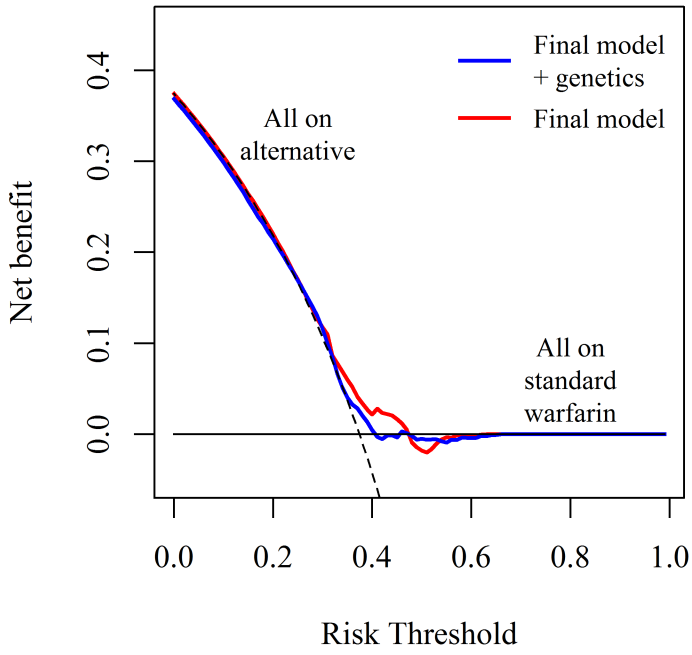

Figure 3.8. Decision curve of prediction models with and without genetic factors. *In this case, the decision curve plots the net benefit of the prediction model compared to the strategy of using standard warfarin treatment in everyone, across the full range of values for the risk threshold. The strategy of using standard warfarin treatment in everyone is shown with the solid black line, while the strategy of using an alternative therapy in everyone is shown with the dashed black line. The net benefit of the strategies of using standard warfarin therapy in everyone and using an alternative therapy in everyone intersects when the risk threshold is equal to the prevalence of the outcome in the overall population. Note that the curve for the all on alternative therapy strategy continues downward beyond the edge of the figure.*

Table 3.5. Characteristics of the derivation and validation cohorts by site.

Variable	HUP[a]		PVAMC[a]	
	Derivation[b] (N = 184)	Validation[b] (N = 263)	Derivation[b] (N = 137)	Validation[b] (N = 198)
Age				
< 45	51 (28)	75 (29)	12 (9)	7 (4)
45 – 55	39 (21)	61 (23)	23 (17)	28 (14)
55 – 65	35 (19)	62 (24)	57 (42)	96 (48)
65 – 75	35 (19)	41 (16)	28 (20)	45 (23)
75+	24 (13)	23 (9)	17 (12)	22 (11)
Female gender	89 (48)	136 (52)	5 (4)	9 (5)
African American race	103 (56)	188 (71)	70 (51)	147 (74)
Body Mass Index				
< 25	63 (34)	66 (25)	41 (30)	62 (31)
25 – 30	62 (34)	77 (30)	36 (26)	57 (29)
> 30	59 (32)	117 (45)	59 (43)	79 (40)
History of hypertension	87 (47)	170 (65)	66 (48)	151 (76)
History of diabetes	40 (22)	68 (26)	50 (36)	71 (36)
History of peptic ulcer disease	16 (9)	26 (10)	17 (12)	9 (5)
History of heart failure	31 (17)	58 (22)	34 (25)	40 (20)
Warfarin indication				
AFib/AFlutter	68 (37)	81 (31)	70 (51)	87 (44)
DVT/PE	73 (40)	131 (50)	40 (29)	86 (44)
Other	43 (23)	51 (19)	27 (20)	24 (12)
Previously used warfarin	47 (26)	75 (29)	40 (29)	72 (36)
Smoking status				
Never	90 (49)	130 (50)	19 (14)	37 (19)
Past	73 (40)	87 (34)	79 (58)	92 (46)
Current	21 (11)	42 (16)	39 (28)	69 (35)
Insurance status				
Private	151 (84)	162 (63)	6 (4)	10 (5)
VA/Medicare/Other	7 (4)	43 (17)	107 (79)	154 (78)
Medicaid/None	21 (12)	54 (21)	23 (17)	34 (17)
Employment status:				
Working	79 (43)	78 (30)	32 (24)	18 (9)
Unemployed	17 (9)	22 (8)	17 (13)	12 (6)
Retired	49 (27)	62 (24)	49 (36)	91 (46)
Disabled	39 (21)	99 (38)	37 (27)	77 (39)
Annual income:				
< $15,000	48 (29)	90 (37)	41 (38)	95 (49)
$15,000 - $20,000	45 (27)	18 (7)	48 (45)	18 (9)
> $20,000	72 (44)	137 (56)	18 (17)	79 (41)

[a]Sites were limited to those that were present in both derivation and validation cohorts.
[b]All values are reported as N (%).

DISCUSSION

Overview. In this study, we sought to develop a model to predict whether a patient starting warfarin would have a prolonged dose-titration phase, and then test the model in an external validation cohort. Given the availability of less burdensome but more expensive alternative oral anticoagulant agents, being able to predict prolonged dose titration could help patients and clinicians decide whether to use warfarin or one of the alternative agents. However, the prediction model we developed failed to validate in an external cohort, with an AUC of 0.59 (95% CI 0.54, 0.64). Thus, our results suggest that it will be difficult for clinicians to predict prolonged dose titration in patients starting warfarin, at least using traditional social, clinical, and genetic predictors.

Model development. The final model contained five variables: warfarin indication, insurance status, number of doctor's visits in the previous year, current smoking status, and history of heart failure. This model performed moderately well in the derivation cohort, with a time-dependent AUC at 12 weeks, as measured by LOOCV, of 0.66 (95% CI 0.60, 0.74). Furthermore, only a moderate amount of shrinkage was needed to improve model calibration, with a linear shrinkage factor of 0.82. Finally, the association between the selected predictor variables and the outcome seemed to be quite stable, as these predictors were seen in the best models across the full range of subset sizes (Figure 3.1).

Model validation. The model performed much worse when tested in the external validation cohort, however. The time-dependent AUC at 12 weeks was only 0.59 (95% CI 0.54, 0.64) when validated externally. Model performance did not improve for the secondary outcomes of reaching maintenance doses within 4 and 8 weeks, indicating that the model's limited ability to discriminate was not unique to a specific time point cut-off. Although the variable selection

algorithm was designed to optimize model discrimination, the model appeared to be reasonably well calibrated in the overall validation cohort, with a non-significant Hosmer-Lemeshow calibration test (P = 0.73). This result suggests that our use of a linear shrinkage factor was largely successful in improving model calibration. However, we believe that model discrimination is the best way to determine the clinical utility of this prediction model, because it would allow clinicians to distinguish between patients at higher risk for prolonged dose titration on warfarin from those at lower risk.

The addition of genetic variants did not improve the performance of the model, with no improvement in the time-dependent AUC observed (P > 0.99). Similarly, the IDI was also poor at 0.01 (95% CI 0.00, 0.02), although it was technically a statistically significant improvement in discrimination. The IDI as a test statistic is known to have problems with type I error, especially as it approaches zero [Kerr, McClelland, Brown, & Lumley, 2011; Pepe, Feng, & Gu, 2008], so this finding of statistical significance should be viewed with skepticism in the context of the rest of our results. Overall, the lack of improvement in prediction from adding genetic factors is consistent with recent clinical trial evidence showing that inclusion of genetic factors in dose prediction models did not lead to significant improvement in clinical outcomes, such as percent time in therapeutic range or time to maintenance dose, over purely clinical dose prediction algorithms [Kimmel et al., 2013].

Differences between derivation and validation cohorts. One reason for the failure of the model to validate is likely the substantial differences between the derivation and validation cohorts, as shown in Table 3.2. Compared with the derivation cohort, the validation cohort was younger, more African American, more obese, more under-insured, and more disabled, among other differences. These differences are likely reflected in the fact that the prevalence of prolonged

dose titration was higher in the validation cohort at 38%, compared to 30% in the derivation cohort. Part of these differences might reflect discrepancies in populations at different sites; for example, the anticoagulation clinic at Johns Hopkins draws from a much more urban African American population than the clinic at Hershey Medical Center. However, there are also substantial differences between the derivation and validation cohorts at the sites that were the same for both cohorts, including the proportion of individuals who are African American, the prevalence of hypertension, the prevalence of different warfarin indications, and the proportion of individuals on disability (Table 3.5). These differences can potentially be attributed to random fluctuations, to changes in the warfarin population or outcomes over time, to differences in practice patterns over time, to differences in those willing to participate, or to changes in recruitment strategies between the two studies. For instance, a decrease in the proportion of patients with atrial fibrillation as their warfarin indication could be related to some of these patients being treated with alternative anticoagulants, which were first approved for that indication. By contrast, the increase in the proportion of patients who were African American at these sites likely reflects recruitment strategies that were designed to increase the enrollment of this group in the validation cohort.

These differences across sites are also reflected by varying performance of the prediction model across sites. For instance, the time-dependent AUC was not significantly better than chance at PVAMC, while it was better at the other sites. Similarly, the model could not account for the substantial differences in baseline risk that was observed at the three sites, with the prevalence of prolonged dose titration varying from 32% to 48%. Moreover, a *post-hoc* analysis where a prediction model was developed and tested using the same algorithm in the sites that were present in both derivation and validation cohorts showed no improvement in model performance (AUC = 0.58; 95% CI 0.51, 0.64), confirming the changes in these same sites over time. Similarly, in

another *post-hoc* analysis, performance of the same model development algorithm in the validation cohort led to the inclusion of some different variables in the model—variables selected were age, BMI, warfarin indication, insurance status, previous warfarin use, history of heart failure, and history of arrhythmia—suggesting that the important predictors of prolonged dose titration might vary across sites. This model did not perform very well on cross-validation (LOOCV AUC = 0.62; 95% CI 0.58, 0.69), suggesting that it also would not perform well on external validation. It should also be emphasized that the broader differences among sites where patients receive warfarin in the clinical community would be expected to be much larger than the differences between our derivation and validation cohort; thus, the performance of the model in clinical practice could be expected to be even worse.

Clinical utility of the prediction model. Attempts to quantify the clinical impact of the prediction model were consistent with our primary results. While the negative predictive value of the model for the lowest range of predicted values (< 20% probability of prolonged dose titration) was reasonably good at 0.87, only 2% of patients in the validation cohort actually fall into this category (Table 3.4). Both positive and negative predictive values were fairly poor at cut-offs that were more commonly observed in our cohort. This drop-off in performance may result from incorrectly ranking individuals in the middle of the probability distribution, which can be seen when plotting the observed vs predicted probabilities by risk decile (Figure 3.4).

As shown in Figure 3.6, the relative utility of the current model is limited, with a maximum value of about 0.09 near the prevalence of the outcome. Additionally, the relative utility is negative for risk thresholds above 0.47, meaning that it is better to use standard warfarin therapy for all patients with high risk thresholds. This impression is confirmed by the related decision curve, which shows that the curves representing the net benefit of the prediction models are not

substantially higher than the curves for the strategies of using the same treatment in everyone for any risk threshold region (Figure 3.8). While the prediction model is unlikely to be useful clinically even in the regions where the relative utility is strictly positive, examination of the risk threshold can still be useful for clinicians. Knowing that the overall prevalence of prolonged dose titration is about 38%, a discussion of the relative importance financial and monitoring burdens with patients can help determine whether treatment with warfarin or an alternative agent is optimal in a given situation. For instance, if a given patient feels that the financial costs of alternative anticoagulant agents are worse than the monitoring burden of warfarin therapy, then his or her individual risk threshold would be above 50%. Since this threshold is greater than the 38% prevalence of prolonged dose titration, it would be optimal to begin standard warfarin therapy in this patient.

Importance of external validation. This study confirms the importance of using external validation when developing clinical risk prediction models. Given its importance, external validation is performed surprisingly infrequently, with recent evidence suggesting that only 25% of published research on new prediction models includes an external validation [Siontis, Tzoulaki, Castaldi, & Ioannidis, 2014]. Especially for complex, multifactorial outcomes like prolonged dose titration for patients starting warfarin, overall prevalence and the importance of different predictors are likely to vary substantially across clinical sites, and even change over time. While statistical methods such as cross-validation can help, external validation remains the gold standard for determining whether a prediction model will be useful in clinical practice.

Conclusions. In conclusion, our prediction model for prolonged dose titration in patients starting warfarin is unlikely to be useful in clinical practice. Moreover, we suspect that this outcome and others like it will be difficult to predict using traditional clinical or genetic risk factors, as their

relationship to the outcome will likely vary substantially across clinical sites. More accurate prediction of prolonged dose titration will likely require researchers to better define and measure the social, behavioral, and access-related factors that are probably more directly related to the outcome. In the absence of risk prediction, clinicians should consider the relative importance of monitoring and financial burdens for their patients when deciding which type of anticoagulation therapy to begin.

CHAPTER 4. IMPROVING CLINICAL PREDICTION MODEL TRANSPORTABILITY WITH SEQUENTIAL UPDATING OF MIXED-EFFECTS MODELS

Brian S Finkelman, Benjamin French, and Stephen E Kimmel

ABSTRACT

Clinical prediction models often fail to generalize across clinical sites outside of those in which the model was derived, and they tend to lose their accuracy over time. These problems have been categorized under the umbrella term of poor model transportability. We propose a general strategy of sequential updating of mixed-effects models as a mechanism to overcome the problem of poor transportability. We examine the potential gains in prediction accuracy for this strategy through a simulation study in which poor transportability is modeled as clinic-specific differences in the prevalence of the outcome and the association between predictors and the outcome. We then test whether the sequential model updating approach is robust to several types of model misspecification.

BACKGROUND

Clinical prediction model transportability. It is well established that clinical prediction models often suffer from the problem of the poor generalizability [König, Malley, Weimar, Diener, & Ziegler, 2007]. In other words, models that perform well in the datasets in which they were derived, measured either by model calibration or discrimination, often perform worse when tested

in other settings. Generalizability of prediction models has been previously described as encompassing two major components: reproducibility and transportability [Justice et al., 1999]. Reproducibility of prediction models can be thought of as the ability of the model to perform well in repeated samples from the same population as the one that yielded the original derivation sample, while transportability refers to the ability of the model to perform well in samples drawn from different but plausibly related populations to the one that yielded the original derivation samples. These plausibly related populations could differ from the original population based on changes over time, geography, clinical setting, and definitions of predictors or outcomes, among other things.

Many statistical methods have been developed to help address the problem of model reproducibility, such as Bayesian model averaging [Hoeting et al., 1999], bootstrap aggregation or bagging [Breiman, 1996], and a variety of methods for cross-validation [Borra & Di Ciaccio, 2010]. Broadly speaking, these methods tend to address the problem of model overfitting. However, poor transportability of a prediction model often occurs because of a problem of underfitting rather than overfitting [Justice et al., 1999]. Underfitting occurs when important predictors are either unknown, misspecified, or not included in the original model, and model performance degrades when tested in new populations with a different conditional prevalence of those predictors. As a result, it is much more difficult to find statistical solutions to problems of transportability using the derivation sample, because by definition, the model would need to be tested on a sample with a different empirical distribution from the derivation sample in order to determine its transportability.

Utility of clinical prediction models is hampered by concerns about poor transportability. Despite the adoption of prediction models in clinical practice, there are often major concerns about model

generalizability and transportability in many clinical scenarios. For instance, the American Heart Association (AHA) and the American College of Cardiology (ACC) released updated cholesterol management guidelines in November 2013 that were heavily based on individuals' predicted 10-year risk of cardiovascular events [Stone et al., 2014]. These guidelines drew almost immediate criticism because of concern about over-prediction of risk related to the particular cohorts used to develop the prediction model [Ridker & Cook, 2013]. Specific examples of validated prediction models failing to generalize to different populations have been documented, as well. For example, the EuroSCORE model, which was developed in European populations to predict 30-day mortality in patients undergoing cardiac surgery, failed to generalize to Australian surgical patients [Yap et al., 2006], and, even with the European population, proved inaccurate over time, over-predicting risk in contemporary practice [Hickey et al., 2013]. In another example, a clinical prediction rule for predicting deep vein thrombosis (DVT) performed well in the secondary referral patient population in which it was developed, but failed to generalize to a primary care setting [Oudega et al., 2005]. Furthermore, this problem is likely even more widespread than what has been directly documented in the literature because of the many clinical outcomes that are known to vary substantially across clinical sites, including readmission after hospitalization for heart failure [Ross et al., 2008], mortality following surgery for colorectal cancer [Schootman et al., 2014], graft failure after liver transplantation [Asrani et al., 2013], and medication adherence rates among diabetes patients [Sherman et al., 2011]. As a result, generally applicable methods to improve the transportability of clinical prediction models could have a large practical impact on a wide range of areas in clinical medicine.

Improving prediction model transportability with sequential model updating. Generalized linear mixed-effects models, also known as longitudinal or hierarchical models, are well-established in the literature for accounting for clustered observations, such as would occur when patients at

specific clinical sites are more similar to each other than to the overall population, in the context of explanatory models [Fitzmaurice, Laird, & Ware, 2011]. However, their utility for improving the transportability of prediction models is less clear, because predictions on novel clusters are based on the hypothetical mean cluster. As a result, there is no heterogeneity across clusters for out-of-sample predictions, even though the model is technically capable of allowing for such heterogeneity. Thus, any improvement in prediction accuracy that results from using mixed-effects models is generally because of shrinkage effects, rather than incorporating knowledge about cluster-specific differences.

One potential approach to the problem of prediction in novel clusters is sequential model updating. Under sequential updating, predictions are made on individuals using the best available model at that time. Then, when their outcome data become available, they are systematically incorporated back into the model. As a result, novel clusters become incorporated into the data sample over time, allowing for predictions that account for cluster-specific differences. In practice, sequential model updating would likely involve incorporating the prediction model into an electronic health records system (EHR) that is integrated across multiple clinical sites, so that outcome data could be automatically captured and incorporated into the model. However, the expected improvement in prediction accuracy that would be achieved through sequential model updating remains unknown. Thus, given the large upfront financial costs and logistical challenges of implementing such a system, it is important to quantify these potential gains, as well as the conditions under which these gains can be maximized.

Simulation study. We sought to quantify the potential improvement in prediction accuracy that might be expected with sequential model updating using a simulation study. Briefly, we simulated a population of patients who are clustered in different clinics. These patients were

randomly split into derivation and validation cohorts. Standard, non-updating prediction models were built in the derivation cohort and then tested in the validation cohort. The same models were then allowed to update periodically to see whether prediction accuracy improved. This process was then repeated 1,000 times to assess the variability of the results. Finally, the sensitivity of the results to changes in the value of parameters for the data-generating process was assessed.

METHODS

Overview. In our simulation, we aimed to develop a model to predict the outcome Y_{ij}, which represents a generic, continuous clinical outcome for patient j at clinic i. Y_{ij} is dependent on X_{1ij}, a known patient-level predictor; X_{2ij}, an unknown patient-level predictor; and N_i, the size of the clinic. Note that X_{1ij} and X_{2ij} can also be interpreted as linear combinations of important predictors, rather than just a single predictor. Clustering of the outcome is induced by a clinic-level random intercept b_{0i} and random slopes b_{1i} and b_{2i}. From 500 total clinics in the population, 20 were randomly selected to make up the "derivation" cohort. Using the derivation cohort, we fit both updating and non-updating versions of models with fixed effects only, as well as those with random intercepts and random slopes. These models were then tested on the remaining clinics, which comprised the "validation" cohort. For each combination of parameter values, the simulation was run 1,000 times to estimate the degree of variability in the results. All simulations were performed using R 3.1.1 [R Development Core Team, 2014].

Mixed-effects modeling. Generalized linear mixed-effects models account for clustering in the outcome by treating some model parameters as random, rather than fixed across the population. These models typically follow the form:

$$g(Y_{ij}) = X\beta + Zb_i + \epsilon_{ij}, \quad (1)$$

where $g(Y_{ij})$ is a function of the outcome for individual j at clinic i, β is a vector of fixed effects, b_i is a vector of random effects, ϵ_{ij} is a vector of residual errors, and X and Z are observed design matrices relating to the fixed and random effects, respectively [Fitzmaurice et al., 2011]. Random effects are typically modeled parametrically as $\mathcal{N}(0, G)$, where G is the variance-covariance matrix. Use of this parametric structure for the random effects is typically more efficient than cluster-level fixed effects, making it especially useful in settings where there are a large number of clinical sites.

Sequential model updating. The primary advantage of combining a sequential model updating approach with generalized mixed-effects models is that it allows the model to automatically calibrate to local conditions, thus improving the transportability of the model, without the need to recruit additional cohorts for constructing and validating a prediction model at each individual site. Additionally, predictions at individual sites are able to "borrow strength" from data at other sites to avoid the overfitting that might occur if separate models were fit at each site. One method of achieving model updating that has been studied in the literature is dynamic logistic regression, in which posterior values for Bayesian model parameters at time *t* are used to construct priors at time *t* + 1, when new data have become available [McCormick, Raftery, Madigan, & Burd, 2012]. However, there are a number of approaches to estimation that could be used to achieve model updating; in this simulation, we are focusing on the simple method of re-fitting the original model at time *t* + 1, after incorporating additional data from the predictions that have been made since time *t*. This choice in estimation allows for a more direct comparison of updating and non-updating models, because all other features of the models are identical.

Data-generating process. For all simulations, we first generated a population of 500 clinics, each with N_i patients, where:

$$N_i \sim \lceil \ln \mathcal{N}(\mu_N, \sigma_N^2) \rceil. \quad (2)$$

The log-normal distribution ensures that there are a large number of smaller clinics, with a small number of very large clinics. The value for μ_N, where $\exp(\mu_N)$ is equivalent to the median clinic size, was fixed at ln(65), while the value for σ_N was fixed at ln(2), in order to ensure a range of clinic sizes of approximately 10 to 500 patients. These values were thought to be reflective of a typical clinical scenario.

Next, clinic-level random intercepts and slopes were generated from a multivariate normal distribution, as follows:

$$\{b_{0i}, b_{1i}, b_{2i}\} \sim \mathcal{N}(0, \mathrm{T}), \quad (3)$$

where b_{0i} is the random intercept, b_{1i} is the random slope for X_{1ij}, and b_{2i} is the random slope for X_{2ij}, and the variance-covariance matrix is:

$$\mathrm{T} = \begin{bmatrix} \tau_0^2 & \rho\tau_1\tau_0 & \rho\tau_2\tau_0 \\ \rho\tau_0\tau_1 & \tau_1^2 & \rho\tau_2\tau_1 \\ \rho\tau_0\tau_2 & \rho\tau_1\tau_2 & \tau_2^2 \end{bmatrix}. \quad (4)$$

The correlation between the random intercept and random slopes, ρ, was fixed at a moderate value of 0.3, which was felt to be similar to what might be observed in practice. However, sensitivity analyses demonstrated that the results were insensitive to increases or decreases in the value of the correlation (data not shown). Additionally, we determined that having the correlation between the random slopes differ from the correlation between the random intercept and random slopes would not have a substantial impact on the results (data not shown), so the same value for all correlations was used for model simplicity. After clinic-level random effects were generated, patient-level variables were generated. First, X_{1ij} and X_{2ij} were generated as $\mathcal{N}(0,1)$ variables. The variance was fixed at 1 for all parameter combinations in order to provide a reference point for easier interpretation of the values of other parameters. We varied τ_0^2 and τ_1^2 in order to determine

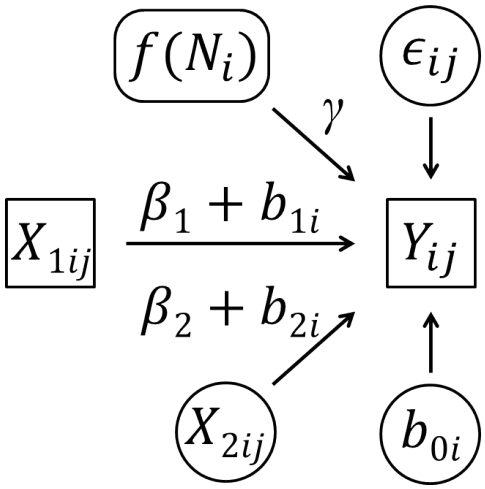

Figure 4.1. Summary of data-generating process. *Variables enclosed in squares are fully observed, variables enclosed in circles are unobserved, and variables enclosed in rounded rectangles are partially observed.*

the impact of different relative strengths of clinic-level heterogeneities, compared to patient-level factors.

Then, the outcome Y_{ij} was generated as:

$$Y_{ij} = b_{0i} + (\beta_1 + b_{1i})X_{1ij} + (\beta_2 + b_{2i})X_{2ij} + \gamma f(N_i) + \epsilon_{ij}, \quad (5)$$

where ϵ_{ij} are independent errors distributed as $\mathcal{N}(0, \sigma_\epsilon^2)$ and the value of σ_ϵ^2 was chosen such that the error terms comprise 20% of the total variance in Y_{ij}. Clinic size is associated with the outcome through the function *f*, with:

$$f(N_i) = \Omega\big(\ln(N_i) - \overline{\ln(N_\iota)}\big), \quad (6)$$

where Ω is a scaling factor such that $f(N_i) \sim \mathcal{N}(0,1)$. The value for β_1 is fixed at 1 across all simulations, so that β_2 and γ gain the interpretation of the impact of X_{2ij} and clinic size on the outcome, respectively, relative to the impact of X_{1ij}. Note that the overall intercept across all clinics, β_0, was defined as equal to 0 and is thus not included in Equation 5. The data-generating process is summarized in Figure 4.1.

Parameter values. The main parameters that were varied for our simulation were τ_0^2 and τ_1^2, which controlled the relative impact of patient-level factors and clinic-level heterogeneities on the outcome. Three values of each parameter were examined—0.5, 1, and 2 for τ_0^2, and 0, 0.25, and 0.5 for τ_1^2—for a total of 9 main parameter combinations. The values of these parameters can be interpreted relative to the size of the variance in X_{1ij}, which was fixed at 1. Additionally, β_2 and γ were fixed at zero for these main parameter combinations, so that the effects of unknown patient-level factors and clinic size on the results could be examined in isolation. When β_2 was equal to zero, τ_2^2 was also set equal to zero, so that there was no effect of X_{2ij} on Y_{ij}; when β_2 was not equal to zero, τ_2^2 was set to be equal to τ_1^2. We considered $\tau_0^2 = 1$, $\tau_1^2 = 0.25$, $\beta_2 = 0$, and $\gamma = 0$ to be the "base" parameter combination, and sensitivity analyses for individual parameters were based on this combination of parameter values.

Later, we separately assessed the impact of non-zero values for β_2 and γ. Specifically, we examined values of $\sqrt{0.5}$, 1, and $\sqrt{2}$ for both parameters. These values were selected for greater interpretability, as the relative contribution of X_2 and $f(N_i)$ to the total variance in Y_{ij} was proportional to β_2^2 and γ^2, respectively. Thus, for example, when $\beta_2 = \sqrt{2}$, X_{2ij} is contributing twice as much to the variance in Y_{ij} as is X_{1ij}. This set of parameter values likely covers the full range of what could reasonably be expected in practice, given that the prediction models were being rigorously constructed in the first place. For this set of parameter combinations, τ_0^2 and τ_1^2 were fixed at their base values.

Finally, we assessed the impact of varying update intervals in an attempt to reflect longer time lags between predictions and the occurrence of the outcome that might take place in certain clinical scenarios, such as those with survival-type outcomes. We examined values of 250, 500, 1,000, and 5,000 for θ, the number of predictions made between rounds of updating for updating

models, as described below. We used θ = 500 as the base value for all previously described parameter combinations.

Prediction models. We randomly selected 20 clinics—stratified by clinic-size quintile, N_i^*—for the derivation cohort, mimicking a multi-site cohort that might be used to develop a clinical prediction model in practice. We selected 6 clinics from each of the bottom 2 quintiles, 3 clinics from each of the next 2 quintiles, and 2 clinics from the upper quintile. We then built 3 prediction models in the derivation cohort:

1) a linear model, $\beta_1 X_{1ij}$;

2) a Bayesian linear mixed-effects (BLME) model, $b_{0i} + \beta_1 X_{1ij}$; and

3) a second BLME model, $b_{0i} + (\beta_1 + b_{1i})X_{1ij}$.

BLME models were fit using restricted maximum likelihood, with non-informative flat priors for the fixed effects and a non-informative prior for the random effects covariance matrix based on the Wishart distribution. Estimation of BLME models was accomplished using the "blme" package in R [Dorie, 2014]. Additionally, for simulations when $\gamma \neq 0$, we also constructed versions of the above models that included N_i^* as a categorical fixed effect, since it was felt that N_i^* would be more likely to be observable than $f(N_i)$ in practice.

All three models were tested in the validation cohort with and without sequential model updating. Sequential model updating was achieved by making predictions on θ patients, incorporating outcome data on those individuals back into the derivation dataset, re-estimating the models, and then making predictions on the next θ patients. This algorithm was repeated until predictions had been made on all patients in the validation cohort. For BLME models, this process was equivalent to adding new data, and did not affect the model priors. The order of predictions was random across the entire validation cohort, and each individual had an 80% chance to have his or her

outcome data incorporated into the updating algorithm. We chose 80% because it realistically allows for missing outcome data; this is reflective of missing outcome data that might occur when utilizing a sequential model updating scheme in practice, where patients might be lost to follow-up before their outcomes are observed.

Assessment of model calibration. Accuracy of prediction models was based on assessments of model calibration, with mean absolute error (MAE) being the primary metric [Wilmott & Matsuura, 2005]. MAE was calculated as:

$$\text{MAE} = \frac{1}{n}\sum\left|\widehat{Y_{ij}} - Y_{ij}\right|, \quad (6)$$

where *n* is the total number of individuals in the validation cohort. To improve the interpretability of the results, we constructed a new metric, the "relative improvement" (RI) in MAE, for each model as:

$$\text{RI} = \frac{\phi_0 - \text{MAE}}{\phi_0 - \phi_1}, \quad (7)$$

where ϕ_0 refers to the mean absolute error for the intercept-only model, as fit in the derivation set, and ϕ_1 refers to the mean absolute error for the "true" model, which was considered to be the model in Equation 4, minus the error term. Thus, the RI will typically range from 0 to 1 and can be interpreted as the improvement of the current model over the intercept-only model, relative to the improvement that would have been seen with the "true" model. Negative values for RI indicate that the given model is worse than predicting the average value in everybody.

RESULTS

Population characteristics. There were 41,576 (SD 1,465) patients in the total simulated population, on average, with 1,276 (SD 118) patients in the derivation cohort. Clinics ranged in

Table 4.1. Clinic size distribution in the simulated population.

Clinic-size quintile[a,b]	0–20%	20–40%	40–60%	60–80%	80–100%
Minimum number of patients in quintile	9 (2)	37 (2)	55 (2)	78 (3)	117 (5)
Percent of population in quintile	6.1 (0.3)	10.9 (0.5)	15.8 (0.6)	23.0 (0.6)	44.2 (1.2)

[a]Results are presented as mean (SD) across 1,000 simulations.
[b]Clinic size distribution is not affected by varying the value of the main parameters.

size from 9 to 549 patients, on average. The median clinic had 66 patients, and 67% of patients were in clinics in the top two quintiles of clinic size. Other characteristics of the distribution of clinic size, which are reflective of the log-normal distribution selected, are shown in Table 4.1. The effect of varying τ_0^2 and τ_1^2 on clinic-level clustering is shown in Figure 4.2; as expected, increasing τ_0^2 , which represents the variance of the random intercepts, tended to yield greater vertical displacement among the slopes, while increasing τ_1^2, which represents the variance of the random slopes, led to a more defined fanning pattern.

Main parameter results. As can be seen in Table 4.2, the prediction models explained a substantial amount of the variance in the derivation cohort, ranging from an r^2 of 0.25 to 0.80, depending on the model and parameter combination. Furthermore, the addition of random effects consistently led to dramatic improvements in the model r^2 in the derivation cohort, creating the initial appearance of improved model performance. However, because all out-of-sample predictions are made assuming that new clinics have the mean value for their random intercept and slope, the addition of random effects led to virtually no improvement in the accuracy of predictions in the validation cohort, with mean RI at 33% to 34% for all non-updating models for the base parameter combination. In contrast, use of sequential model updating led to dramatic improvements in RI for both BLME models, across all parameter combinations tested (Figure 4.3).

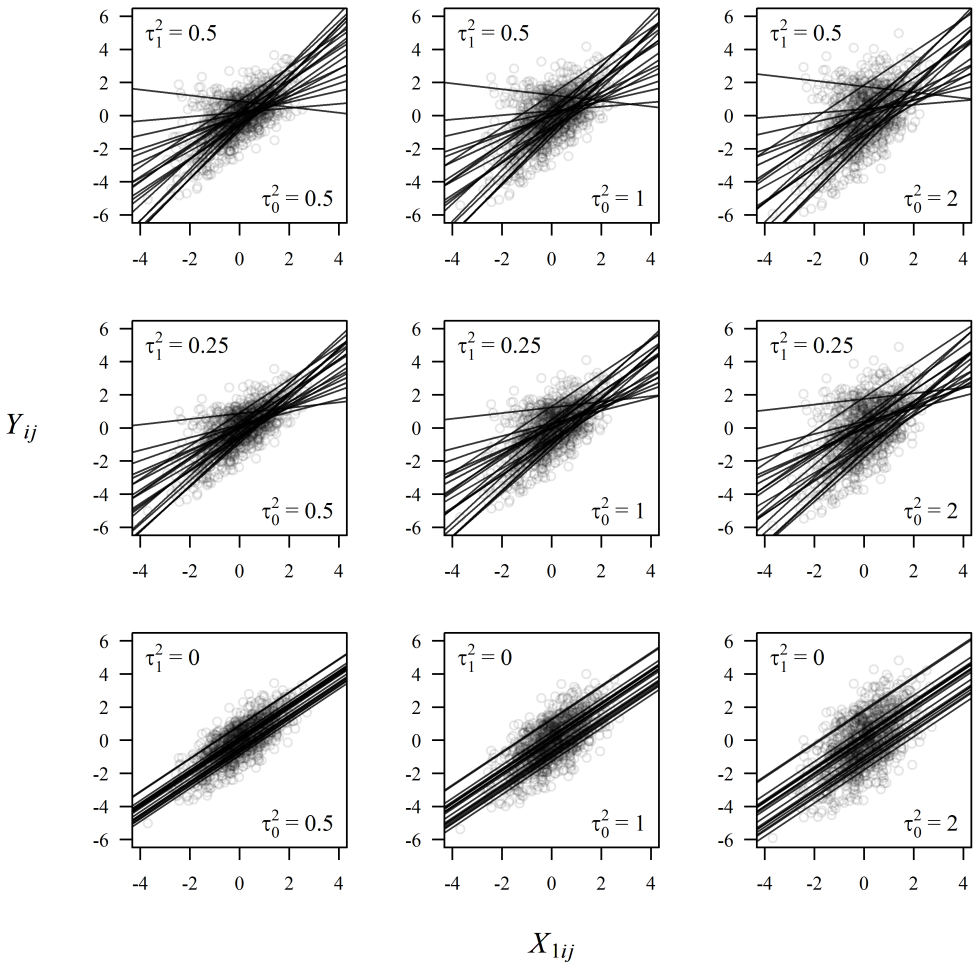

Figure 4.2. Effect of τ_0^2 and τ_1^2 on clinic-level clustering. *Each point represents an individual patient at one of the 20 clinics in the derivation cohort for a single simulation run. Lines represent the actual relationship between X_{1ij} and Y_{ij} at each derivation clinic. The center figure represents the base parameter combination.*

Table 4.2. Mean r^2 for non-updating models in derivation cohort across all main parameter combinations.

τ_1^2	Model	$\tau_0^2 = 0.5^a$	$\tau_0^2 = 1^a$	$\tau_0^2 = 2^a$
0.5	$\beta_1 X_{1ij}$	0.42 (0.11)	0.34 (0.10)	0.25 (0.09)
	$b_{0i} + \beta_1 X_{1ij}$	0.61 (0.08)	0.64 (0.07)	0.68 (0.07)
	$b_{0i} + (\beta_1 + b_{1i}) X_{1ij}$	0.80 (0.04)	0.80 (0.04)	0.79 (0.04)
0.25	$\beta_1 X_{1ij}$	0.47 (0.09)	0.38 (0.08)	0.27 (0.08)
	$b_{0i} + \beta_1 X_{1ij}$	0.69 (0.06)	0.71 (0.05)	0.73 (0.06)
	$b_{0i} + (\beta_1 + b_{1i}) X_{1ij}$	0.80 (0.04)	0.80 (0.04)	0.79 (0.04)
0	$\beta_1 X_{1ij}$	0.55 (0.06)	0.42 (0.06)	0.29 (0.06)
	$b_{0i} + \beta_1 X_{1ij}$	0.80 (0.02)	0.79 (0.03)	0.79 (0.04)
	$b_{0i} + (\beta_1 + b_{1i}) X_{1ij}$	0.80 (0.02)	0.80 (0.03)	0.79 (0.04)

[a]Results presented as mean (SD) over 1,000 simulations.

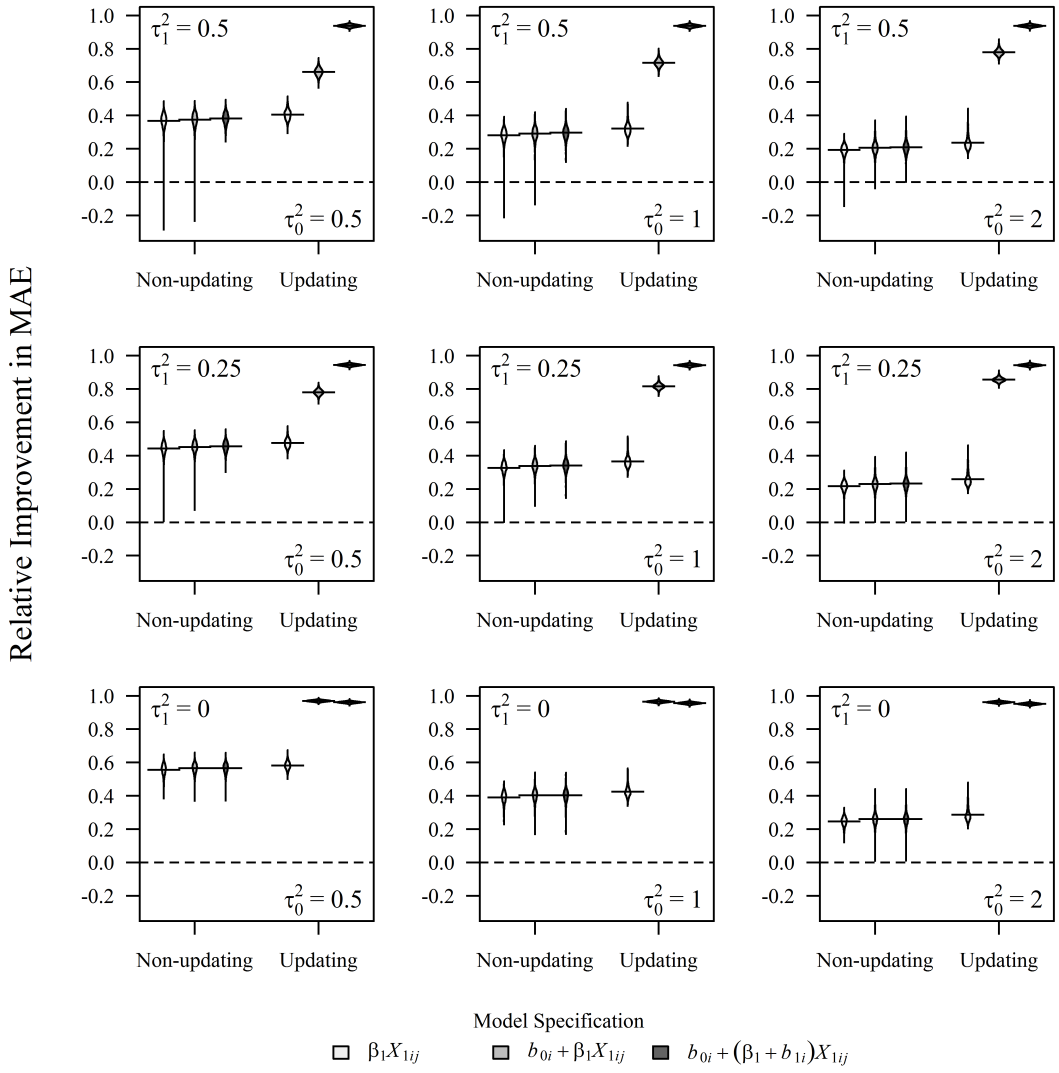

Figure 4.3. Relative improvement in MAE for both updating and non-updating models across all main parameter combinations. *Plots show the density of values for relative improvement in MAE across 1,000 simulations, with horizontal bars representing the mean value.*

As can be seen in Figure 4.4, gains in prediction accuracy from sequential model updating were seen across all clinic-size quintiles, although the greatest improvement was seen in the largest clinics. This pattern likely reflects the fact that improvements from updating were seen relatively rapidly, with approximately 90% of the total gains in predictive performance for both BLME

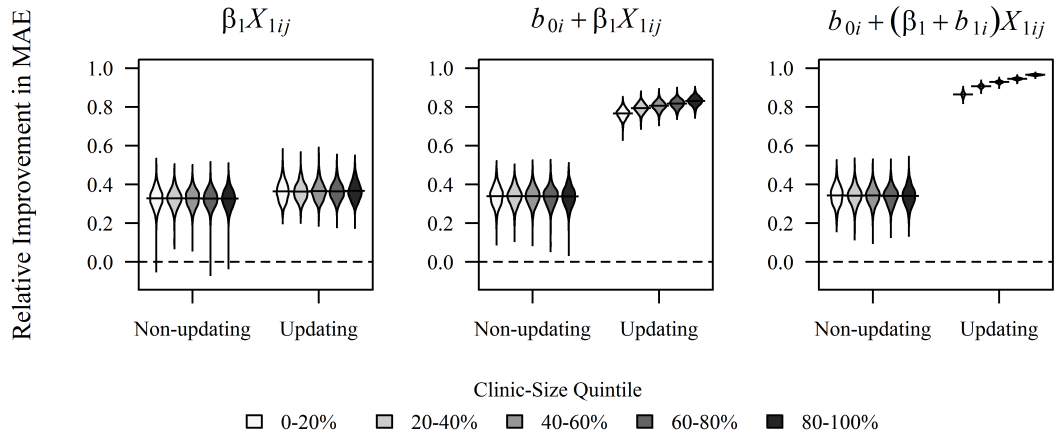

Figure 4.4. Relative improvement in MAE by clinic-size quintile. *Plots show the density of values for relative improvement in MAE across 1,000 simulations, with horizontal bars representing the mean value. These results are for the base parameter combination.*

models occurring after about 10 predictions at a given clinic (Figure 4.5). Because there were 480 clinics in the validation cohort and the model was updated after every 500 predictions, model updates occurred after almost every prediction, on average, especially at smaller clinics.

Effect of model misspecification. When there was an unknown patient-level factor impacting the outcome (i.e. $\beta_2 \neq 0$), sequential model updating was less effective (Figure 4.6). However, updating models still were more accurate than non-updating models for all values of β_2. Having the outcome be dependent on clinic size (i.e. $\gamma \neq 0$) led to worse performance of non-updating BLME models, with these models performing worse than intercept-only models with large values of γ (Figure 4.7). However, updating BLME models showed no drop-off in prediction accuracy with non-zero values of γ. Inclusion of clinic size quintile, N_i^*, as a categorical fixed effect led to marked improvement in non-updating BLME models and even slight improvement in updating BLME models, on average (Figure 4.8).

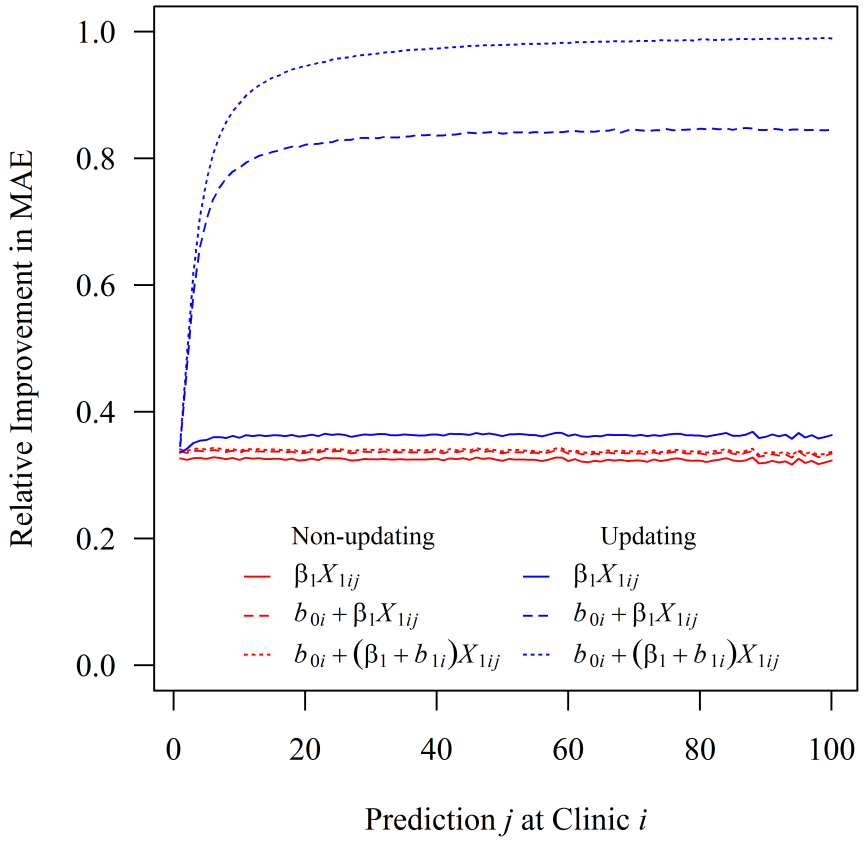

Figure 4.5. Rate of improvement in prediction accuracy at a given clinic. *This plot shows the mean relative improvement in MAE for prediction j at clinic i, across 1,000 simulations. These results are for the base parameter combination.*

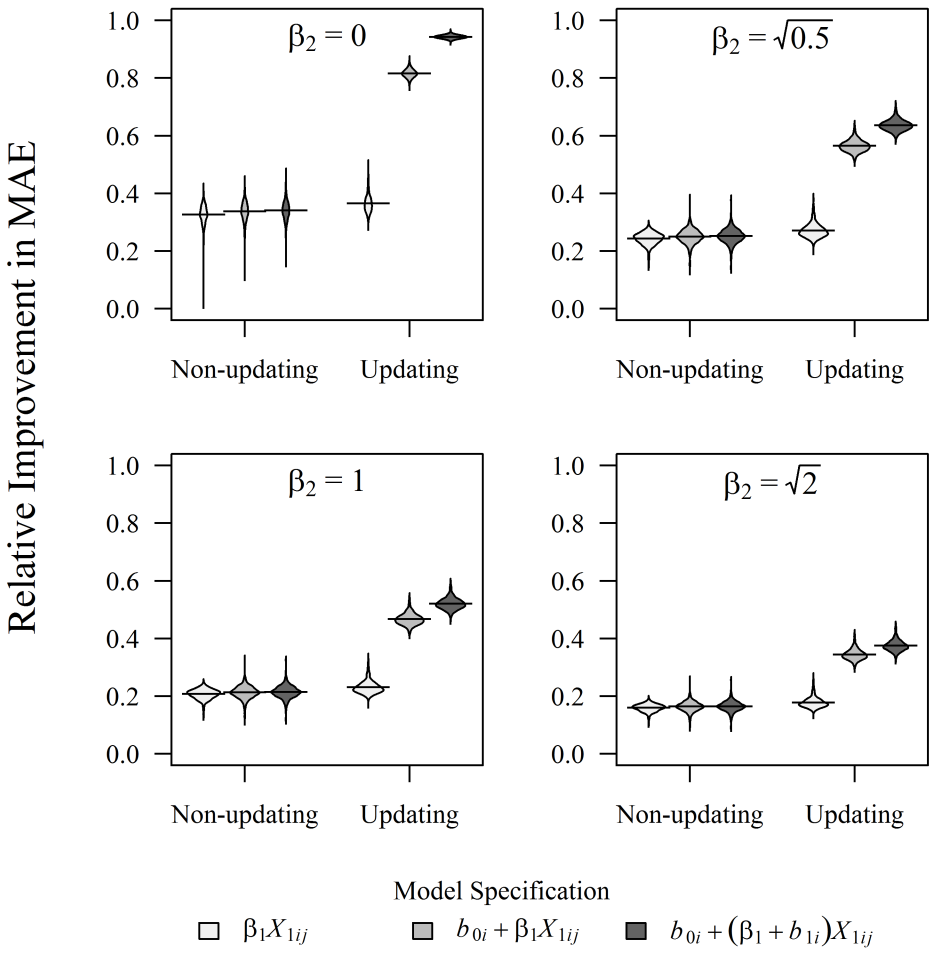

Figure 4.6. Effect of β_2 on model prediction accuracy. *Plots show the density of values for relative improvement in MAE across 1,000 simulations, with horizontal bars representing the mean value. The parameters for τ_0^2 and τ_1^2 are fixed at their base values. Note that the relative contribution of X_{2ij} to the total variance in Y_{ij}, compared to X_{1ij}, is equal to β_2^2.*

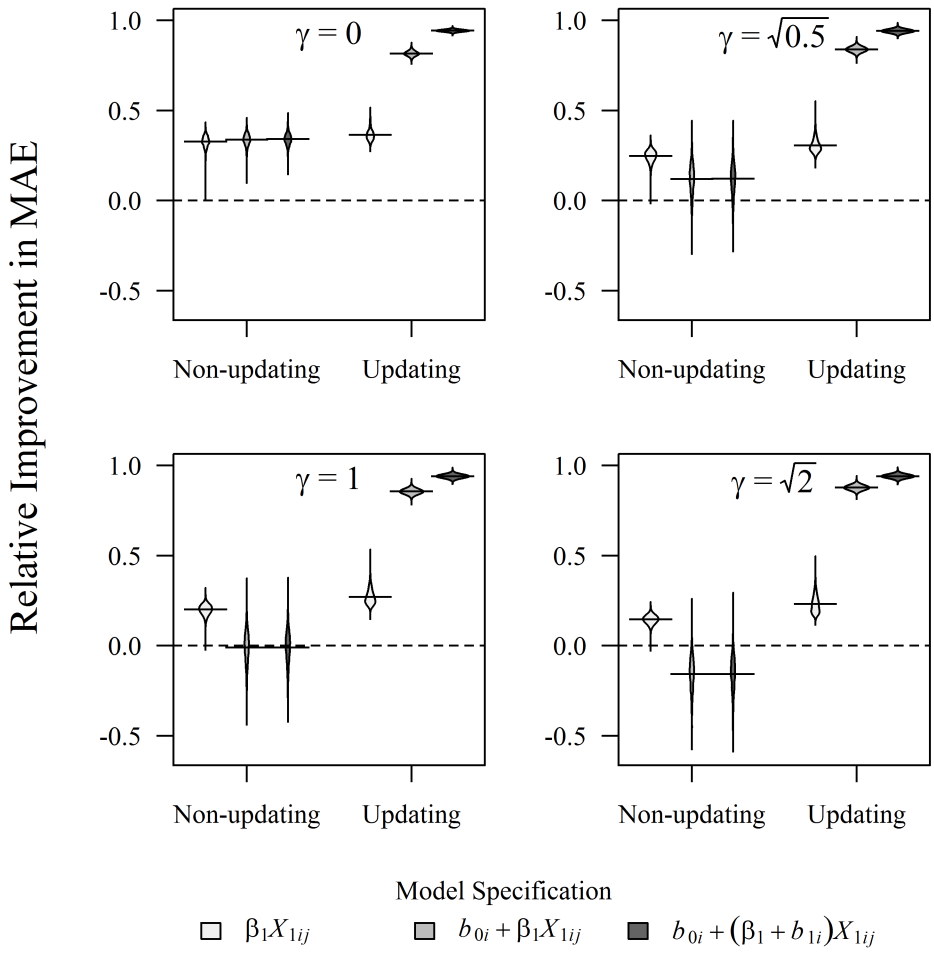

Figure 4.7. Effect of γ on model prediction accuracy. *Plots show the density of values for relative improvement in MAE across 1,000 simulations, with horizontal bars representing the mean value. The parameters for τ_0^2 and τ_1^2 are fixed at their base values. Note that the relative contribution of $f(N_i)$ to the total variance in Y_{ij}, compared to X_{1ij}, is equal to γ^2.*

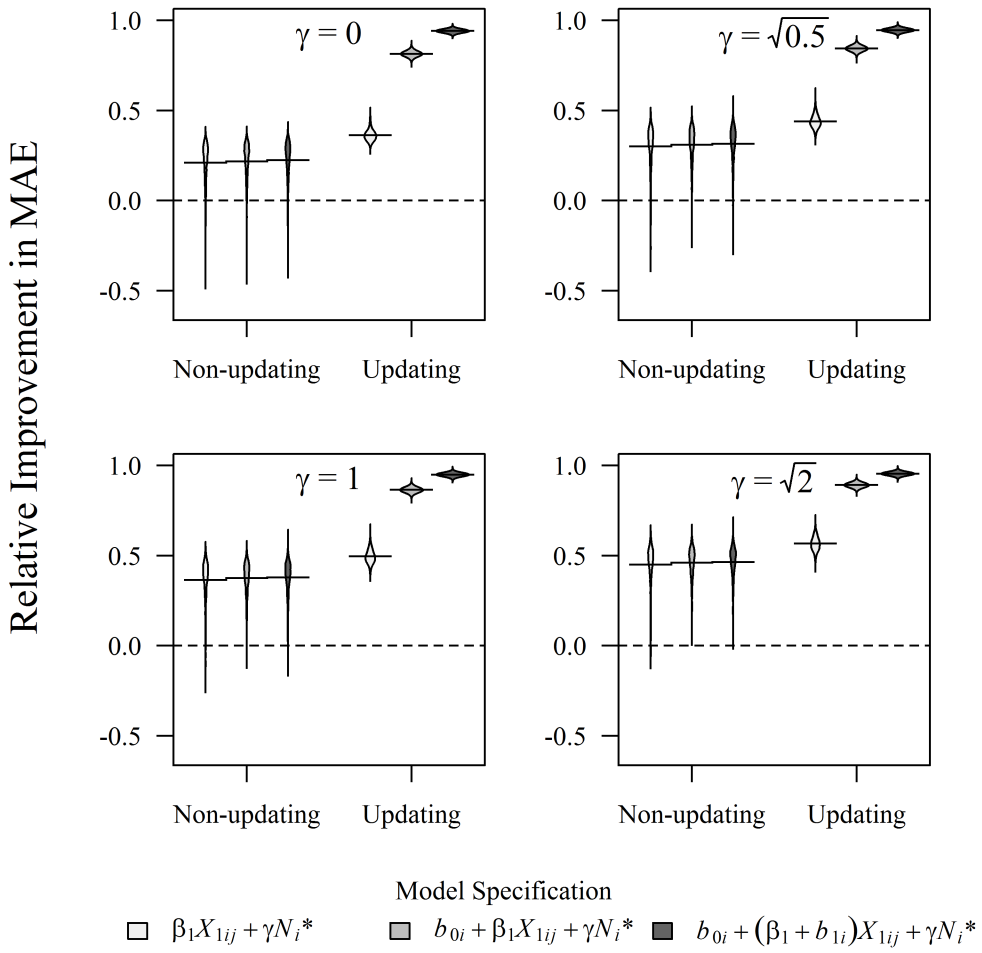

Figure 4.8. Effect of γ on prediction accuracy for models that include N_i^*. *Plots show the density of values for relative improvement in MAE across 1,000 simulations, with horizontal bars representing the mean value. All models include N_i^*, which represents clinic-size quintile, as a categorical fixed effect. The parameters for τ_0^2 and τ_1^2 are fixed at their base values. Note that the relative contribution of $f(N_i)$ to the total variance in Y, compared to X_1, is equal to γ^2.*

Effect of varying the update interval. Results were fairly insensitive to changes in θ, the update interval. Even when θ = 5,000, or about 12.5% of the validation cohort, prediction accuracy in updating BLME models was not substantially decreased (Figure 4.9). Furthermore, prediction accuracy was consistent across all quintiles of clinic size with varying values of θ (data not shown). Finally, the rate of improvement in prediction accuracy only showed a notable decrease when θ = 5,000, when about 90% of total gains in prediction accuracy for both BLME models occurred after about 20 predictions at a given clinic (Figure 4.10). Note that this value for θ corresponds to a highly unlikely scenario where the model can only be updated about 8 times over the course of using the model on a population of about 40,000 individuals.

DISCUSSION

Overview. In this simulation study, we sought to quantify the potential effect of sequential model updating on the accuracy of clinical prediction models. Sequential updating of BLME models led to uniform improvement in prediction accuracy across all parameter combinations examined. Thus, it seems quite likely that substantial gains in the transportability of clinical prediction models could be achieved through sequential updating of models that account for clinic-specific heterogeneities, including differences in the mean level of the outcome as well as differences in the association between known predictors and the outcome. However, the extent of the gains in prediction accuracy from updating varied depending on the degree of misspecification of fixed effects, indicating that use of sequential model updating will likely be more useful in clinical scenarios where such misspecification can be minimized.

Impact of sequential model updating. Sequential model updating did not substantially improve prediction accuracy with the linear model, performing similarly to all non-updating models.

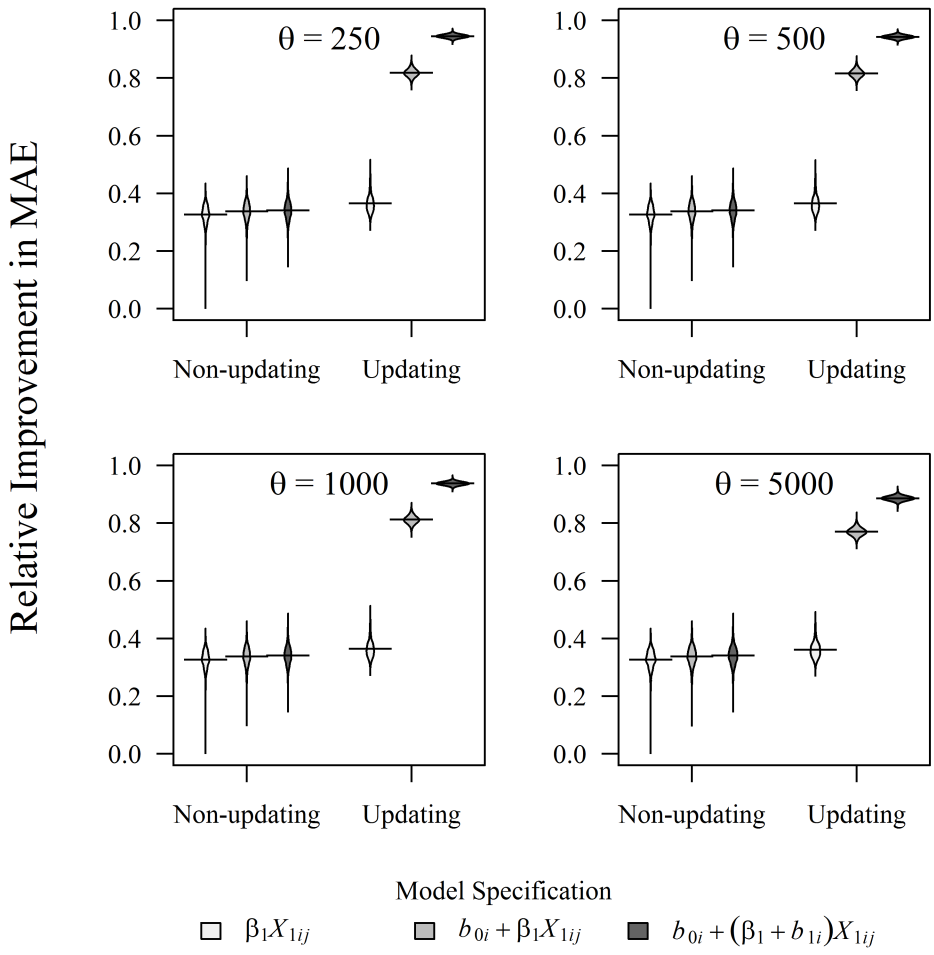

Figure 4.9. Effect of the update interval, θ, on model prediction accuracy. *Plots show the density of values for relative improvement in MAE across 1,000 simulations, with horizontal bars representing the mean value. All other parameters are fixed at their base values.*

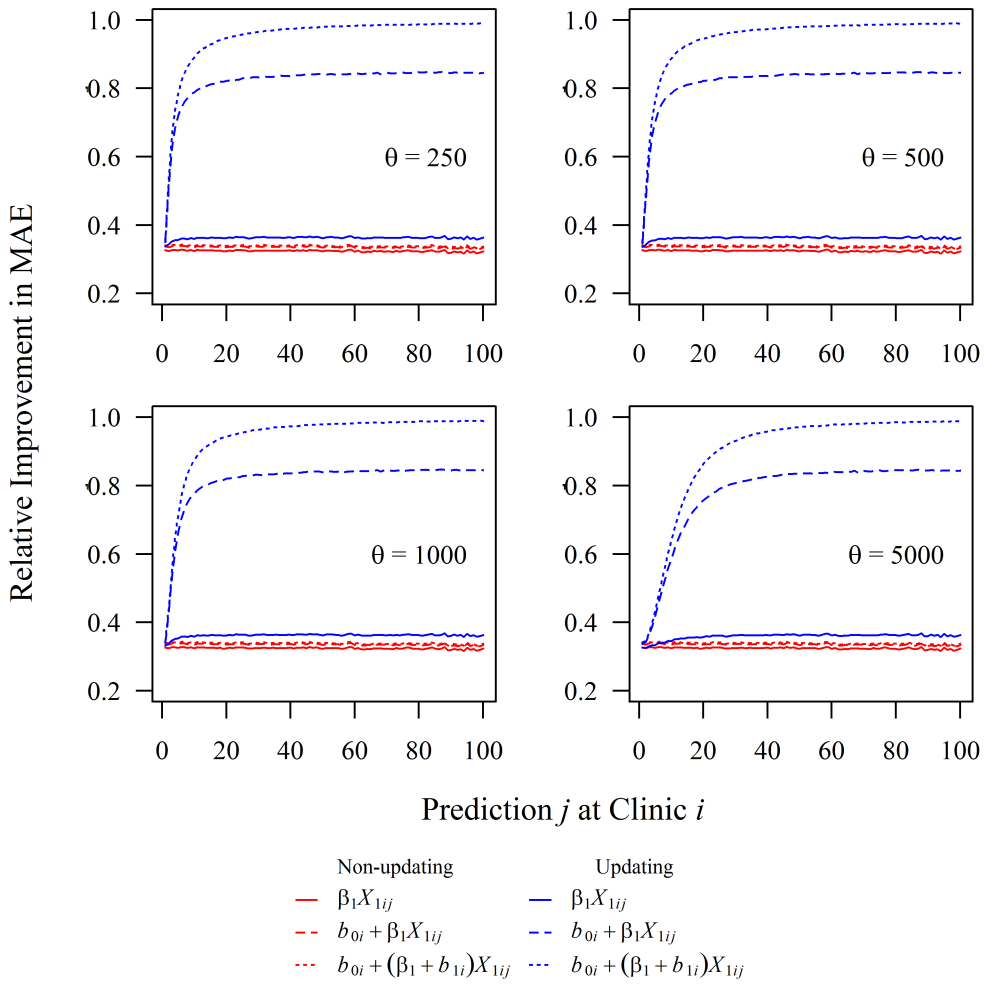

Figure 4.10. Effect of the update interval, θ, on the rate of improvement in prediction accuracy at a given clinic. *This plot shows the mean relative improvement in MAE for prediction j at clinic i, across 1,000 simulations, for different values of* θ. *All other parameters are fixed at their base values.*

As a result, flexible models with random effects were needed to account for the heterogeneities across clinics. The accuracy of non-updating models decreased with increasing values of τ_0^2 and τ_1^2; in short, greater heterogeneity across clinics led to worse performance for models that did not take these differences into account. By contrast, updating BLME models showed greater improvement in prediction accuracy when a larger proportion of the variation in the outcome was explained by clinic-level heterogeneities. The BLME model with a random intercept showed improved prediction accuracy with increasing values of τ_0^2; however, its performance deteriorated with higher values of τ_1^2. This deterioration in accuracy with larger random slopes is not surprising, because this model had no way to account for the random slopes that were present in the data structure. Even so, the model was able to use its random intercept to account for a large enough amount of inter-clinic variability to provide uniform improvement over non-updating models and the linear updating model.

The BLME model with both a random intercept and random slope was nearly as accurate as the "true" model across all main parameter combinations, with a mean RI ranging from 94 to 96%. This was because the model was essentially equivalent to the data-generating model in these cases, and updating occurred fast enough that predictions on most individuals in the validation cohort were made with a fully calibrated model. Indeed, about 90% of the gains in prediction accuracy were seen by about the 10^{th} patient at a given clinic, so even small clinics were able to see benefits from sequential model updating, and the majority of predictions at large clinics were made with an accurate estimate of clinic-specific random effects. This rapid improvement in prediction accuracy was largely sustained even with higher values of θ, so overall prediction accuracy in the validation cohort was preserved even when models were updated less frequently. It should also be noted that this high level of prediction accuracy was sustained even when there was no random slope in the data-generating process ($\tau_1^2 = 0$). Thus, there was not really much

downside to having an unnecessary random slope in the updating model, while having only a random intercept when the data-generating process included both a random intercept and a random slope led to decreased prediction accuracy.

Additionally, the variance of RI values across simulations tended to be lower in updating models than in non-updating models. The variance in prediction accuracy decreased with each additional random effect in the model, as well. This speaks to another important feature of sequential model updating, which is the ability to overcome sampling bias to produce models that perform more consistently. In the non-updating models, the prediction accuracy was largely dependent on whether the clinics that comprised the derivation cohort happened to be representative of the overall population. In simulations where estimates of β_0 and β_1 were very different from their true values due to random sampling, prediction accuracy for non-updating models in the validation cohort tended to be worse (Figures 4.11–4.13). However, sequentially updating models were able to overcome initial sampling bias by improving model calibration over time.

Impact of model misspecification. When an unknown patient-level factor was added to the data structure ($\beta_2 \neq 0$), updating BLME models had a decrease in prediction accuracy; however, they still performed better than non-updating models for all values of β_2. In short, it is still important to be diligent when selecting covariates and their specifications for a sequentially updating model, as models that are closest to the true data-generating process will still perform the best. However, most realistic clinical scenarios involve unknown predictors and misspecification, so the fact that sequential model updating still led to improvements in prediction accuracy under these conditions suggests that it may be a useful strategy in the real world.

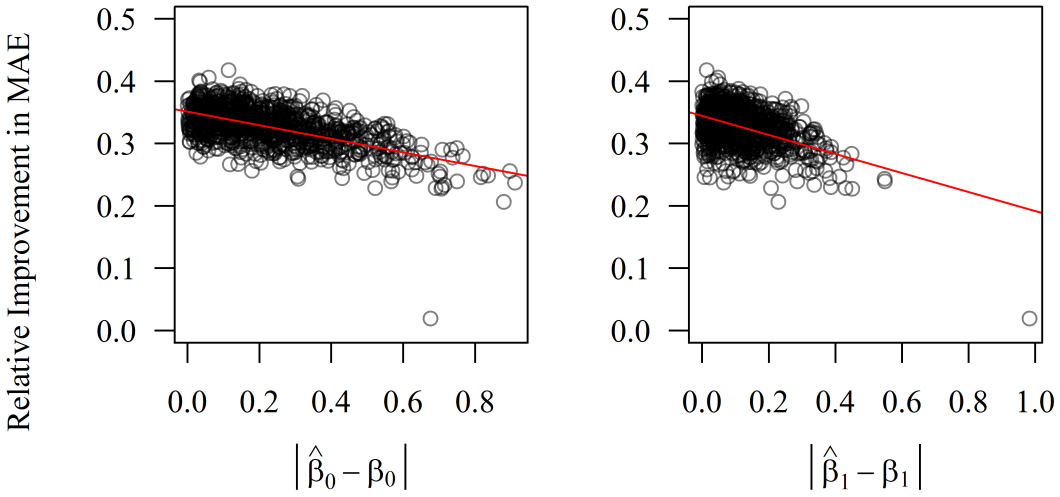

Figure 4.11. Relationship between bias in estimated model coefficients and prediction accuracy for the linear model. *Each point represents one of 1,000 total simulations for the base parameter combination, and best fit lines are shown in red. The left panel shows the bias in the estimated intercept from the derivation cohort compared to the true value in the overall population, while the right panel shows this bias for the estimated slope.*

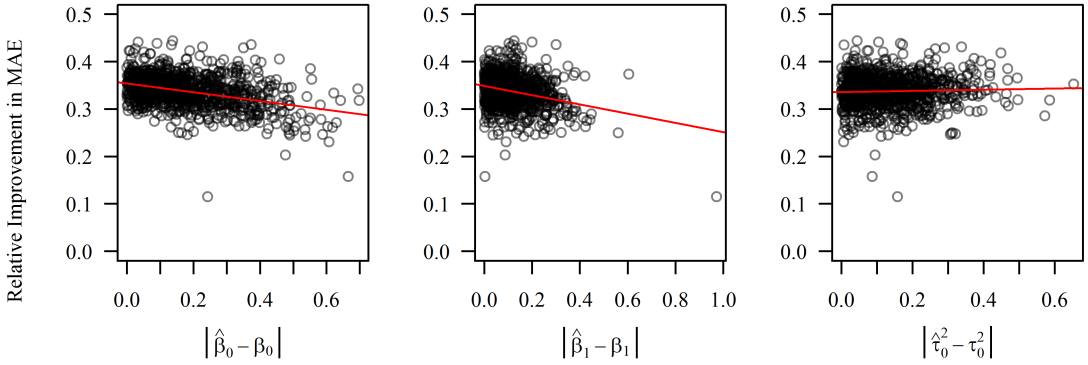

Figure 4.12. Relationship between bias in estimated model coefficients and prediction accuracy for the BLME model with random intercept. *Each point represents one of 1,000 total simulations for the base parameter combination, and best fit lines are shown in red. The left panel shows the bias in the estimated intercept from the derivation cohort compared to the true value in the overall population, the middle panel shows this bias for the estimated slope, and the right panel shows this bias for the estimated variance of the random intercepts.*

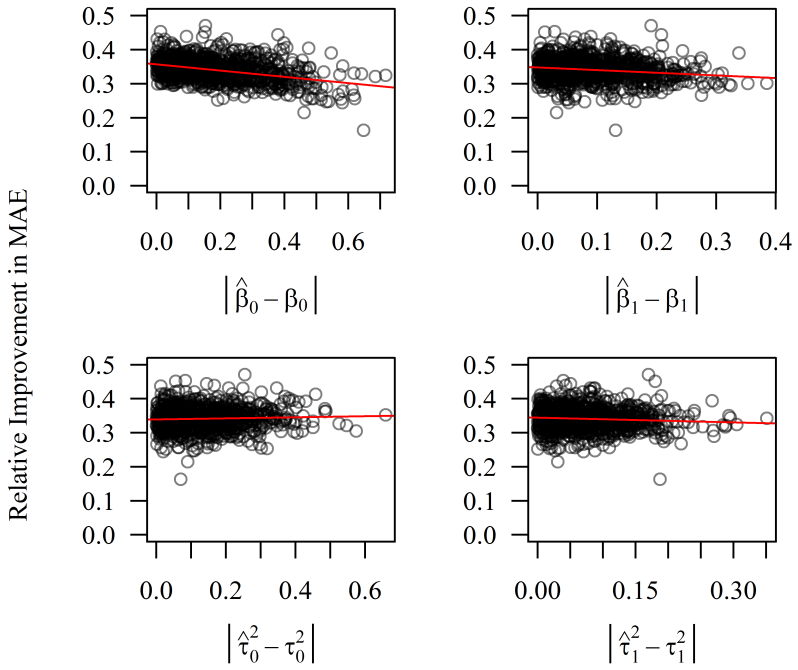

Figure 4.13 Relationship between bias in estimated model coefficients and prediction accuracy for the BLME model with random intercept and slope. *Each point represents one of 1,000 total simulations for the base parameter combination, and best fit lines are shown in red. Starting from the top left panel and moving in clockwise fashion, the panels show the bias in the estimated intercept, slope, variance in the random slopes, and variance in the random intercepts, as compared to the true value in the overall population.*

Clinic size or volume may be related to outcomes in a number of clinical scenarios, such as hospital mortality rates for acute myocardial infarction or surgical mortality rates [Birkmeyer et al., 2002; Silber et al., 2010]. While other clinic-level effects can be accommodated by random intercepts and slopes, we were concerned that clinic size might behave differently because it is directly related to the probability of observing the data in the first place. Larger values of γ led to worse performance of non-updating BLME models, while updating BLME models showed no deterioration in performance. In non-updating BLME models, the effect of sampling bias was actually amplified because differences due to clinic size were incorporated into the model as random effects, with greater bias in the estimated random effects covariance matrix leading to

worse prediction accuracy (Figures 4.14 and 4.15). However, in updating models, these initial biases were diminished over time because the model was continually being calibrated to the overall population, such that the majority of predictions were unaffected by the initial biases. As a result, inclusion of N_i^* was required to improve the accuracy of non-updating BLME models, but not practically necessary in the case of updating BLME models. These results also speak to the general robustness of sequentially updating models that account for clinic heterogeneities; while correct specification is still better, misspecification is not nearly as costly as it is with non-updating models.

Challenges to incorporating sequential model updating in practice. Implementation of sequential model updating in practice will likely involve many logistical and analytical challenges. In order to work well, prediction models will likely need to be integrated into EHR systems, so they will be able to automatically extract covariate data to make an initial prediction, and then automatically extract outcome data to use for model updating. Furthermore, in order to accommodate heterogeneities across sites, the EHR will need to either be standardized across all of the sites, or compatible enough to allow for communication of data. Additionally, the data storage and security requirements for large amounts of data across multiple sites will likely be quite complex. Certain analytic strategies—such as Bayesian dynamic regression, where posterior distributions are estimated from dynamic priors in a fully online fashion [McCormick et al., 2012]—could greatly reduce the data storage requirements, and, accordingly, the data security concerns. However, more work is needed to determine the trade-offs in prediction accuracy that might accompany this approach under certain scenarios. Finally, there will need to be a concerted effort to communicate the effectiveness of this approach to the clinical community in order to foster the necessary level of trust to overcome initial financial and logistical hurdles.

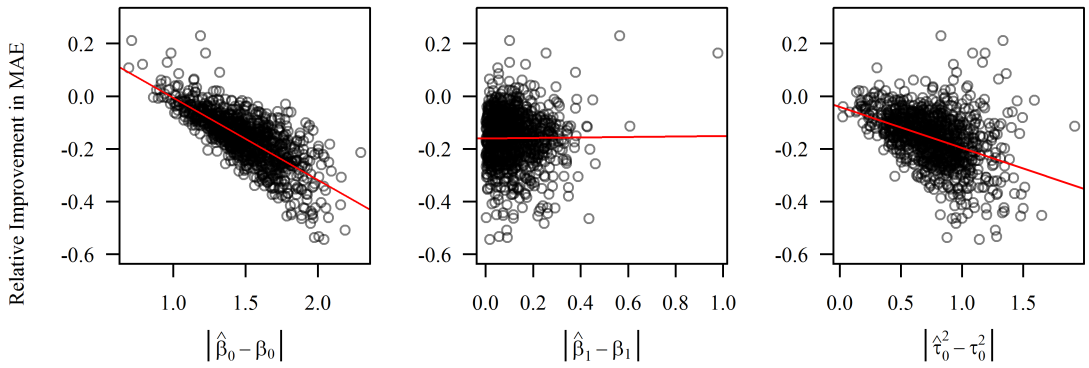

Figure 4.14. Relationship between bias in estimated model coefficients and prediction accuracy for the BLME model with random intercept, with clinic size influencing the outcome. *Each point represents one of 1,000 total simulations with* $= \sqrt{2}$*, and best fit lines are shown in red. The left panel shows the bias in the estimated intercept from the derivation cohort compared to the true value in the overall population, the middle panel shows this bias for the estimated slope, and the right panel shows this bias for the estimated variance of the random intercepts.*

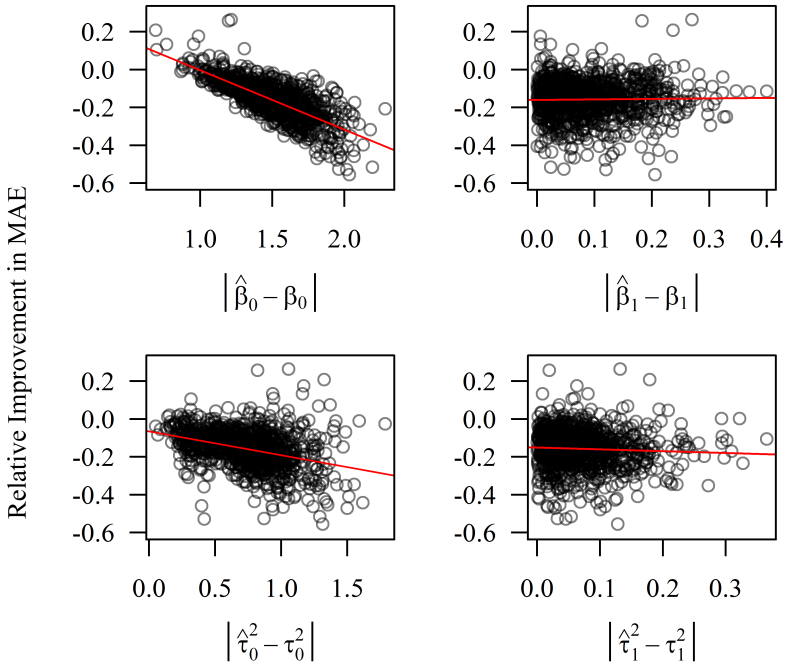

Figure 4.15 Relationship between bias in estimated model coefficients and prediction accuracy for the BLME model with random intercept and slope, with clinic size influencing the outcome. *Each point represents one of 1,000 total simulations with* $= \sqrt{2}$*, and best fit lines are shown in red. Starting from the top left panel and moving in clockwise fashion, the panels show the bias in the estimated intercept, slope, variance in the random slopes, and variance in the random intercepts, as compared to the true value in the overall population.*

The analytic challenges involved in sequential model updating are also likely to be quite complex. Missing data, both for covariates and outcomes, will be an important issue to resolve, as standard methods, such as multiple imputation [Groenwold, Donders, Roes, Harrell, & Moons, 2012; Moons, Donders, Stijnen, & Harrell, 2006], may be difficult to implement in the context of a dynamic system. As a result, efforts to jointly model the updating process along with the prediction model itself, analogous to methods for jointly modeling longitudinal and competing risks data [Li, Elashoff, & Li, 2009], may be required. Alternatively, use of missing indicators may be more useful than with standard models [van der Heijden, Donders, Stijnen, & Moons, 2006], because these parameters would be allowed to calibrate to the population over time. However, further studies are needed to answer these questions empirically. Other important analytic issues that would need to be resolved include how to incorporate new predictors or specifications into a sequentially updating model; how much to weight historical data in a population that is changing over time; how best to account for time lags between making predictions and obtaining outcome data; and how to determine whether a model is not performing well enough at a given site and needs to be replaced with a separate, newly derived model.

Study limitations. Although our simulation was based on a hypothetical predictor and outcome variable, we tried wherever possible to mimic situations that might occur when developing and utilizing a typical clinical prediction model. For instance, we utilized a log-normal distribution for clinic size, so that there would be a larger number of small clinics than large clinics, and we generated the derivation cohort to be similar in size and composition to what might be found in a large multi-center cohort study. We also excluded some patients from contributing data to updating models, to reflect the loss to follow-up that might occur in clinical practice. Finally, we examined scenarios where the model was not correctly specified, which are likely to occur in real-world applications.

Despite these efforts, there are a number of limitations to our model. For instance, we did not examine scenarios where heterogeneities across clinics were not normally distributed. It is possible that standard BLME models might not perform as well in this scenario, leading to a model that was less calibrated to local conditions, even after updating. However, research studying the impact of misspecified parameterization of random effects on prediction accuracy suggests that the standard multivariate normal assumptions should be reasonably robust [McCulloch & Neuhaus, 2011]. Additionally, we assumed in our simulation that outcome data that was not available for updating was missing completely at random, which may not hold in practice. Future studies are needed to determine whether the prediction accuracy of sequentially updating models will be worsened in scenarios where the probability of obtaining outcome data for updating is dependent on model covariates or, especially, the outcome.

We attempted to cover a reasonable range of parameter values in our analysis; however, it is possible that our results will not extrapolate to values outside of the tested ranges. Additionally, to reduce computational burdens, we focused on a simplistic model: a single continuous predictor and a continuous outcome. Clearly, clinical prediction models in the real world will have multiple covariates, and many will have more complex outcomes. The precise gains in prediction accuracy from sequential model updating will likely vary depending on the particular structure of the data in question. Finally, sequential model updating in practice will have to deal with a lag between when predictions are made and when outcomes are observed. It is possible that long lag periods relative to the frequency of updating will decrease the rate at which prediction accuracy improves. As a result, sequential model updating may be less useful for prediction models with long lag times, especially at smaller clinics or in rapidly changing populations. We attempted to assess the sensitivity of our results to long lag times by varying the update interval, θ, and large improvements in prediction accuracy with updating BLME models were still seen even at the

highest values of θ. Even with these positive results, though, the exact effect of time lags on the performance of sequentially updating models will need to be formally addressed in future research. Despite these limitations, we are confident that sequential model updating will prove to be a useful approach for a broad set of clinical scenarios.

Future directions. Many of the limitations and challenges discussed above provide an excellent framework for future research in this area. More simulations are needed to test the performance of sequential model updating in the context of clustered populations that change dynamically over time, which will be more reflective of actual patient populations. Additionally, more rigorous study of time lags in outcome variables and approaches to handling missing data are needed. Furthermore, more explicit comparison of different specific modeling approaches, such as formal Bayesian dynamic approaches [McCormick et al., 2012], model averaging techniques [Raftery, Gneiting, Balabdaoui, & Polakowski, 2005], non-parametric strategies [Ryu, Li, & Mallick, 2011], and machine learning methods [Hastie et al., 2009] are needed. Finally, these approaches will need to be tested in a variety of empirical studies to determine the extent to which theoretical gains are likely to be realized in practice.

Conclusions. In conclusion, sequential updating of models that accommodated clinic-level differences led to improved prediction accuracy in the overall population. The extent of the improvement in prediction accuracy that was observed with updating mixed-effects models depended on the relative impact of clinic-level and patient-level factors on the outcome as well as the degree of model misspecification; however, updating mixed-effects models were uniformly superior to non-updating models as well as updating models with only patient-level fixed effects. Gains in prediction accuracy tended to occur rapidly, leading to improvements at small clinics as well as large clinics. While there are many logistical and analytical questions to resolve, the

potential for a sequential model updating approach to improve the transportability of clinical prediction models seems quite promising.

CHAPTER 5. CONCLUSIONS

In Chapter 2, we found that baseline behavioral factors, health care utilization, and health quality were associated with longer time to maintenance dose in patients initiating warfarin therapy, while in Chapter 3, we discovered the difficulty of developing a model to predict prolonged dose titration in these patients. Our results in Chapter 4 suggested that sequential model updating of mixed-effects models can lead to substantial improvement in prediction model transportability. In addition to these specific results, however, a major focus of this dissertation was using warfarin response as an example of therapeutic effectiveness research in general. Thus, while the studies in Chapters 2 and 3 are designed to address specific questions about patients initiating warfarin therapy, the lessons gleaned from these studies can apply to the field of therapeutic effectiveness research more broadly. Similarly, the methods examined in Chapter 4 would be expected to extend beyond models of therapeutic effectiveness to clinical prediction models more generally.

In Chapter 2, we examined the genetic, clinical, and social factors associated with time to maintenance dose (TTM) for patients starting warfarin therapy. The results highlight the importance of considering non-genetic factors when studying outcomes related to anticoagulation control. While most of the previous research on TTM had focused on genetic variants that have been previously found to affect the required therapeutic dose of warfarin [Cavallari et al., 2011; Higashi et al., 2002; Jorgensen et al., 2009; Limdi et al., 2008; Meckley et al., 2008], none of the genetic variants we examined were significantly associated with TTM. Instead, TTM appeared to be more related to baseline behavioral factors, health care utilization, and health quality. Of particular importance was the finding that having previously been on warfarin was associated with longer, rather than shorter, TTM. This new finding suggests that clinicians should be just as vigilant in monitoring these patients, even though they have more experience with warfarin

therapy. These findings are salient for the broader field of therapeutic effectiveness research, as well. In an era of "personalized medicine" [Crews, Hicks, Pui, Relling, & Evans, 2012], it is important that research on the impact of genetic factors on the effectiveness of a given therapy not come at the expense of research on non-genetic factors, which may be just as important, if not more so, in clinical practice.

When conducting this study, we had hoped to identify potential targets for future interventions for improving TTM in patients on warfarin therapy by examining the effect of post-initiation factors, such as changes in interacting medications or changes in diet. Our results were disappointing here, because none of the post-initiation factors examined were significantly associated with TTM. These results further suggested that changes likely did not occur frequently enough in the early stages of warfarin therapy to affect TTM or that clinicians responded to these changes with appropriate dose changes. However, it is still possible that these factors might be more important for determining anticoagulation control in patients in the maintenance phase of therapy, when monitoring is less frequent and dose titration is not active. Future research on this topic will need to ensure correct specification of time-varying factors to avoid immortal time bias as well as adjustment for variable INR monitoring frequency to avoid interval censoring bias.

In Chapter 3, we developed and externally validated a model to predict prolonged dose titration in patients initiating warfarin therapy. While the model developed appeared to perform well in the derivation cohort, even when assessed using cross-validation, it did not perform as well in the external validation cohort. As a result, it is unlikely that the model will be useful in clinical practice. *Post-hoc* analyses suggested that model performance varied substantially across clinical sites, with marked differences in the AUC among the anticoagulation clinics in the validation cohort. These sites differed from each other in terms of outcome prevalence and patient

characteristics, and the most important predictors of prolonged dose titration in the validation cohort were somewhat different from the derivation cohort. Although the prediction model itself will not be useful for clinical practice, our results offer an important cautionary tale on the essential need for external validation when developing prediction models.

Furthermore, the rigorous decision-theoretic approach that we used to examine the clinical utility of our model will still be useful, both to clinicians and to future researchers. An understanding of the risk threshold can help clinicians formally think about the relative costs of financial and monitoring burdens for their patients and then come to a decision about optimal treatment choice based on the overall prevalence of the outcome. Additionally, future prediction models on therapeutic effectiveness will likely be more easily incorporated into clinical practice if they can demonstrate their usefulness to the clinical decision-making process with metrics such as relative utility. More research is certainly needed to develop summary metrics of prediction model performance that are rooted in decision theory; it is likely that clinicians will be more trusting of these methods when they become simpler and more intuitive.

The substantial variability in the performance of our prediction model across clinical sites in Chapter 3 provided an unexpectedly good motivation for the methodological work done for the project described in Chapter 4. Poor transportability is a pervasive problem for clinical prediction models, and, generally, most research has focused on developing new models or finding new predictors that can provide incremental improvement, without addressing the fundamental challenge of accounting for variability in the relationship between predictor variables and outcomes in different locations, across clinical domains, and over time. With recent technological advancement in and widespread adoption of electronic health record (EHR) systems [Kukafka et al., 2007], it has become easier to imagine systems that utilize EHR data to improve predictions

made across integrated health systems. Essentially, prediction models could be incorporated into an EHR system, used to make predictions on patients within that system, and then updated sequentially as outcome data on those patients become available. Flexible models, such as mixed-effects models, would thus be able to use this updated information to calibrate to local conditions, such as individual clinics or even individual patients in some settings, over time, while using the data from the overall population to avoid overfitting at any one site.

Because integrating a clinical prediction model into an EHR would likely involve substantial upfront costs, we felt it was important to quantify the potential gains in prediction accuracy that could be achieved by sequential model updating. We achieved this aim through a simulation study, presented in Chapter 4, comparing the prediction accuracy of several updating and non-updating models for a generic clinical outcome. The results suggested that sequential updating of models that account for heterogeneity across clinics in mean outcome levels and predictor-outcome associations can lead to dramatic improvements in prediction accuracy. Furthermore, while the extent of the gains varied depending on the degree of model misspecification—including misspecification of the random effects structure, the presence of unknown patient-level predictors, and the presence of unknown or misspecified clinic-level predictors, such as clinic size—there were no scenarios we examined in which updating models performed worse than non-updating models. Thus, sequential model updating has the potential to be a broadly applicable approach to improving clinical prediction modeling.

However, there remains important methodological work to be done before sequential model updating approaches can be widely adopted in clinical practice. For instance, the length of time between when predictions are made and when outcomes are experienced could impact the feasibility of sequential model updating in certain clinical scenarios. Additionally, methods to

deal with missing data and outcome-dependent data collection will need to be tested. Different types of statistical models will also need to be compared based on how well they perform in an updating framework. Metrics to decide how to incorporate new predictors into established models and to determine whether stratified or unified models are needed across diverse patient populations will need to be developed. Finally, empirical demonstration projects are also likely to reveal unanticipated logistical and analytic challenges that can form the basis for future methodological research. All of this work will help to clarify the types of clinical situations where sequential model updating would be expected to be most useful and how best to implement this approach in practice.

Ultimately, adaptation is likely to be a common theme for therapeutic effectiveness research moving forward. Anticoagulation research is shifting in focus from how to determine a patient's warfarin dose to how best to use warfarin as one of a number of therapeutic alternatives. Although it proved to be less useful for predicting prolonged dose titration in patients starting warfarin therapy, genetic and genomic data will likely be more successful at predicting who is likely to respond to therapy or experience side effects for other specific conditions. In contrast, methods to improve medication access and adherence will likely be more important for conditions where genetic factors are less useful. To maximize their clinical utility, prediction models will need to be able to adapt to heterogeneities in patient populations and practice patterns in different locations as well as changes in clinical practice over time. It is our hope that the work in this dissertation and the work that will arise from it represent a small step toward making therapeutic effectiveness research more effective.

APPENDIX

List of medications considered to interact with warfarin. Potentially interacting drugs were identified from the Physicians Desk Reference, Drug Facts and Comparisons 4.0, and MEDLINE literature searches as of the time that patients were enrolled in the study.

Drug Name	Drug Name (cont.)
(CHOLESTROL LOWERING MED.) LIPITOR	IMURAN
ASTORVASTATIN	INDOCIN (X 7 DAYS)
ATORVAST	ISONAL
ATORVASTATIN	KETOCONAZOLE
ATORVASTATIN CA/LIPITOR	KETOCONAZOLE (PILLS)
ATORVASTATIN CALCIUM	KETOCONAZOLE CREAM
ATORVASTIN CALCIUM	KOFECOXIB
ATOVASTITIN CALCIUM	LAMISIL
CRESTOR	LASIX
CRESTOR/ROSAVASTATIN	LASIX INCREASED
HIGH CHOL. MED./ LIPITOR	LEVAGUIN
LESCOL	LEVAQUIN
LESCOL 20MG QD	LEVAQUIN 750
LESCOL/FLUVASTATIN	LEVAQUIN/ANTIBIOTIC
LIPITOL	LEVOFLOXACIN
LIPITOR	LEVOFLOXCIN
LIPITOR 20MG GD	LEVOQUIN
LIPITOR 40MG QD	LEVOQUIN (TILL 10-16)
LIPITOR/ATORVASTATIN CA	LEVOTHROYOXINE
LIPITOR/ATORVASTATIN CALCIUM	LEVOTHYROPINE
LIPOTON	LEVOTHYROXIN
LOSCOL	LEVOTHYROXINE
LOVASTANTIN	LEXA PRO/SELECTIVE SERETONIN REUPTAKE
LOVASTATIN	LISINIPRIL/HCTZ/ZESTORETIC
MEVACOR	LISINOPRIL/HYDROCHLROTHIAZIDE/ZESTORETIC
PRAVACAL	MASOCORT AC / NASAL STEROID
PRAVACHOL	MAXIDE
PRAVACHOL 80 MG DAILY	MEDROL
PRAVACHOL/PRAVASTATIN	METHIMAZOLE
PRAVACHOT	METHIMAZOLE THYROID
PRAVASTATIN/PRAVACOL	METHONIDAZOLE
PRAVOCHOL	METHYLPHENIDATE

PREVASTATIN	METHYL-PREDINZONE
PROVOCHOL/PRAVASTATIN	METOLAZONE
ROSUVASTATIN/CRESTOR	METRINIAZOLE
SIMAVASTATIN	METRONIDAZOLE
SIMVASTATIN	METRONIDAZOLE X 8 WKS.
SIMVASTATIN (ZOCOR)	MOTRIN
SIMVASTIN	MULTI-SYMPTOM NON-ASPIRIN COLD MEDICINE
ZOCOR	NAFCILLIN
ZOCOR 20 MG PO QD	NAPROSIN
ZOCOR/SIMVASTATIN	NAPROSYN
ZOCOR/SIMVASTIN	NAPROXEN
ZOLCOR	NAPROXIN
(CANCER TREATMENT) "CARBOPLATIN	NAPROXYN
(CHEMOTHERAPY) S-FU	NASACORT
(HYDROCODONE-APAP)	NASOCORE
(PAIN MED.) HYDROCO/APAP	NASOCORT
(TERBINAFINE)	NELFINAVIR
A.S.A.	NELFINAVIRMESYLATE/NIRACEPT
A.S.A. 81	NEOMYCIN
ACARBASE	NIZOVAL CREAM
ACCOLATE	NORVIR
ACCURETIC	OLMESARTAN MEDOXOMIL/HCTZ/BENICAR
ACETAMINOPHEN	OMACOR (FISH OIL) RX
ACETAMINOPHIN	OMAPRAZOLE
ADVIL	OMEGA 3 FATTY ACIDS
ADVIL COLD PILLS	OMENPRAZOLE
ALDACTAZIDE/SPIRONOLACTONE	OMEPRAZOLE
ALDACTONE	OMEPRAZOLE (GERD)
ALEVE	OMEPRAZOLE/PRILOSEC
ALFALFA	OMEPROZOLE
ALKA SELTZER PLUS COLD MED.	OMESARTAN/HYDROCHLOROTHIAZIDE
ALLAPURINOL	ORTHOTRYCYCLINE
ALLIPURINOL	OXALIPPATIN
ALLOPURINOL	OXYCODONE W/APAP
ALLUNOPURINOL	OXYCODONE/APAP
ALLUPROPINOL	PANADOL FOR COLD
AMIADARONE	PARACETAMOL
AMIODARON	PAROXETINE
AMIODARONE	PAXIL
AMIODARONE HCL	PCE
AMIODARONE HCL.	PENICILLIN (PENECILLIN)
AMIODIONE	PERCOCET
AMIODORONE	PERCOCET (OXYCODONE-APAP
ANTIBIOTIC/METRONIDAZOLE	PERCOCET PRN
ANTIBIOTICS-	PERCOCET-POSTOP PAIN

SULFAMETHOXOZOLE/TRIMETHOPRIM	
ANTIDEPRESSANT-CELEXA	PERCOCETS
APA	PERIOSTAT
APA/TYLENOL	PESTO-CET (PERCOCET)
ARTHROTEK	PHENOBARBITAL
ASA	PHENYTOIN
ASA (FOR PROCEDURE)	PHENYTOIN SODIUM
ASA 81 MG PO QD	PHYTONADIONE
ASA 81MG.	PIROXICAM
ASA, 81MG	PIROXICAM/FELDENE
ASIPRIN	PIROXICAM/FELDINE
ASPIRIN	PIROXICAN
ASPIRIN (LOW DOSE)	PIROXICN
ASPRIN	PLACIDEL
AZATHIOPRINE	PREDNISOLONE EYEDROPS
AZATHIOPRINE/IMURAN	PREDNISONE
AZITHROMYCIN	PREDNIZONE
AZITHROMYCIN-ONE DOSE ONLY	PREDUIBONE
AZMACORT	PREDUISONE
BABY ASA	PRILOSEC
BACTRIM	PRILOSEC 11/29-12/12/04
BACTRIM SS/SULFAMETHOXAZOLE/TRIMETHOPRIM	PRIMIDONE
BACTRIM/SULFAMETHOXAZOLE/TRIMETHOPRIM	PRIOXICAM
BACTRUM	PROPAFANONE
BENICAR	PROPAFENONE
BEXTRA	PROPAFENONE (RYTHMOL)
BEXTRA 10MG QD	PROPAFENONE-RYTHMOL
BEXTRA/VALDECOXIB	PROPAFERONE
BEXTRA/VALDECOXILO	PROPANOLOL
BIAXIN/CLARITHROMYCIN	PROPANOLOL ER
BICALURIMINE	PROPOXYPHENE.
BICALUTAMIDE	PROPRANOLOL
CAPECITABINE	PROTOZONE
CARAFATE	PROXICAM/FELDENE
CARBOPLATIN	PROZAC
CASODEX	QUININE
CASODEX/BICALUTAMIDE	QUININE SULFATE
CEFAZOLIN	RANITIDINE
CEFRIAXONE	RANITIDINE HCL
CEFTRIAXONE	RANITIDUIE
CELEBREX	RANTITIDEINE
CELEBREX 200MG	REFOCOXIB
CELEBREX/ CELECOXIB	REQUIP
CELECOXIB/CELEBREX	REQUIP (RLS)/ROPINIROLE HCL

CELEXA	REQUIP/REPINIROLE
CHEMOTHERAPY-TAXOL	REYATAZ
CHLORPHENIRAMINE MALEATE	REYATAZ (ATAZANAVIT)
CHOLESTYR	RHYTHMOL
CIPRO	RHYTHMOL 300MG TID
CIPRO 3/29 -> 4/2/05	RIBAURIN
CIPROFLOXACIN	RIBAVIRIN
CITALOPRAM	RIFAMPIN
CITALOPRAM HYDROBROMIDE	RITALIN
CLARITHROMYCIN	RITONAVIR (NORVIR)
CLELBREX	RYTHMOL
CLOBETASAL CREAM	SANDOSTATIN
COATED ASPIRIN	SERTRALINE
CONCERTA (PRN)	SPIRONALACTONE
CORTISONE SHOT	SPIRONOLACTONE
CORTIZONE SHOT	SPIRONOLACTONE/HCTZ
CYCLOSPORIN	SULFA
DARVOCET	SULFAMETHOXAZOLE
DARVOCET-N	SULINDAC
DECADRON	SUSTIVA
DEPAKOTE (BIPOLAR)	SYNTHROID
DETROL	SYNTHROID/LEVOTHYROXINE
DETROL-LA	SYNTHROID-1 MG.
DEXAMETHASONE	TAXOL
DEXAMETHASONE/DECADRON	TEQUIN/GATIFLOXACIN
DEXAMETHAZONE	TERBINAFINE HCL
DILANTIN	TERBINAFINE/LAMISIL
DILANTIN/PHENYTOIN	TEROZASIN
DIURETIC LASIX	TESTOSTERONE (ANDRODERM PATCH)
DOXERCALCIFEROL (FOR PARATHYROID)	TESVOSVERONE
DOXYCYCLINE	TETRACYCLINE
DOXYCYLINE	THALIDOMIDE
ECOTRIN	THERAFLU
EFUDEX	TOLTERODINE TARTRATE/DETROL
ENDOCET	TOOK 1ST NAPROSYN-"
ENDOCET/PERCOCET	TOOK 2ND RELAFEN-"
ENDOCOT (STOOL SOFTENER)	TOPAMAX (MIGRANES)
ENSURE	TRAMADOL
ERYTHROMYCIN	TRAMADOL/CENTRAL ANALGESIC
ERYTHROPOIETIN	TRAZAD
ERYTHROPOIETIN/EPOGEN	TRAZADONE
ERYTHROPOIETIN-EPOGEN	TRAZADONE HCL
ESTRACE	TRAZODONE
ETODOLAC/FOR PAIN	TRENTAL 400MG PO TID
ETOPOSIDE	TRIAMCINOLONE
EXCEDRIN TENSION HEADACHE	TRIAMCINOLONE CREAM

EXTRA STRENGTH TYLENOL	TRIAMCINOTONE CREAM
FLAGYL	TRICOR
FLAGYL/METRONIDAZOLE	TRICOR/FENOFIBBRATE
FLORAZEMIDE	TYLENOL
FLUDROCORTISONE ACETATE	TYLENOL 3
FLUOXETINE	TYLENOL 500
FLUROSEMIDE	TYLENOL COLD
FLUROSIMIDE	TYLENOL COLD & SINUS
FRESH FROZEN PLASMA	TYLENOL PM
FUOROSEMIDE	TYLENOL PM (PRN)
FUROSEMIDE	TYLENOL PRN.
FUROSEMIDE / DUIRETIC	TYLENOL SINUS
FUROSEMIDE 40MG DAILY	TYLENOL W/CODEINE
FUROSEMIDE/DIURETIC	TYLENOL WITH CODEINE
FUROSEMIDE/DUIRECTIC	TYLENOL/CODENE
FUROSEMIDE/DUIRETIC	TYLOX
FUROSEMIDE/H20 PILL	ULTRACET
FUROSEMIDE/LASIX	ULTRACET MCN 2 EVERY 4-6 HR. AS NEEDED
FUROSEMIDE-DUIRETIC	ULTRAM
FUROSIMIDE/H2O PILL	VALPROIC ACID
GATIFLOXACIN	VICODAN
GEMFIBROZIL	VIOX
GENERIC PERCOCET	VIOXX
GLUCOSAMINE/CHONDROITIN	VIRACEPT
H2O PILL - LASIX (GENERIC)	VIT C
H2O PILL-HYDROCHLOROTHIAZIDE	VIT E
H2O PILLS/FUROSEMIDE	VITAMIN C
HALOPERIDOL	VITAMIN E
HCLT	VITAMIN K.
HCT2	VYTORIN
HCTZ	XELODA
HCTZ (DIURETIC)	XELODA (XELODA)
HCTZ (HYDROCHLOROTHIAZIDE)	ZANTAC
HCTZ/HYDROCHLOROTHIAZIDE	ZAROXALYN (METOLAZONE)
HCTZ/TRIAMTERENE	ZITHRO PAC
HTCL	ZITHROMAX
HTCZ	ZITHROMYCIN
HYDR0CHLOROTHIAZIDE	ZOLOFT
HYDROCHLOROTHIAZIDE	ZOLOFT/SERTRALINE
HYDROCHLOROTHIAZIDE (HCTZ)	ZOSYN
HYDROCHLORTHESIDE/HCTZ	ZOSYN ONE DOSE
HYDROCHLORTHIA ZIOLE	Z-PAC
HYDROCHLORTHIAZIDE	Z-PAC/ZITHROMAX
HYDROCHLORTHIZIDE	Z-PACK
HYDROCHLOTHIAZID	Z-PACK ANTIBIOTIC (TOOK 9-23 TO 9-28)
HYDROCO/APAP	Z-PACK-5 DA ONLY

HYDROCORTISONE	Z-PAK/ZITHROMAX
HYZAAR	ZYRTEC/CETIRIZINE
IBUPROFEN	

BIBLIOGRAPHY

Agnelli, G, & Becattini, C. (2008). Treatment of DVT: how long is enough and how do you predict recurrence. *J Thromb Thrombolysis*, 25(1): 37–44.

Akritas, MG. (1994). Nearest Neighbor Estimation of a Bivariate Distribution Under Random Censoring. *Ann Stat*, 22(3): 1299–1327.

Ansell, J. (2010). Warfarin versus new agents: interpreting the data. *Hematol Am Soc Hematol Educ Progr*, 2010: 221–228.

Apostolakis, S, Sullivan, RM, Olshansky, B, & Lip, GYH. (2013). Factors affecting quality of anticoagulation control among patients with atrial fibrillation on warfarin: the SAMe-TT2R2 score. *Chest*, 144(5): 1555–1563.

Arnsten, JH, Gelfand, JM, & Singer, DE. (1997). Determinants of compliance with anticoagulation: A case-control study. *Am J Med*, 103(1): 11–17.

Asrani, SK, Kim, WR, Edwards, EB, et al. (2013). Impact of the center on graft failure after liver transplantation. *Liver Transpl*, 19(9): 957–964.

Avorn, J. (2011). The relative cost-effectiveness of anticoagulants: obvious, except for the cost and the effectiveness. *Circulation*, 123(22): 2519–2521.

Baker, SG. (2009). Putting risk prediction in perspective: relative utility curves. *J Natl Cancer Inst*, 101(22): 1538–1542.

Baker, SG, Cook, NR, & Vickers, A. (2009). Using relative utility curves to evaluate risk prediction. *J R Stat Soc*, 172(4): 729–748.

Beyth, RJ, Quinn, LM, & Landefeld, CS. (1998). Prospective evaluation of an index for predicting the risk of major bleeding in outpatients treated with warfarin. *Am J Med*, 105(2): 91–99.

Birkmeyer, JD, Siewers, AE, Finlayson, EVA, et al. (2002). Hospital volume and surgical mortality in the United States. *N Engl J Med*, 346(15): 1128–1137.

Borra, S, & Di Ciaccio, A. (2010). Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. *Comput Stat Data Anal*, 54(12): 2976–2989.

Bosworth, HB, Granger, BB, Mendys, P, et al. (2011). Medication adherence: a call for action. *Am Heart J*, 162(3): 412–424.

Breiman, L. (1996). Bagging predictors. *Mach Learn*, 24(2): 123–140.

Brotman, DJ, Jaffer, AK, Hurbanek, JG, & Morra, N. (2004). Warfarin prophylaxis and venous thromboembolism in the first 5 days following hip and knee arthroplasty. *Thromb Haemost*, 92(5): 1012–1017.

Bushnell, CD, Olson, DM, Zhao, X, et al. (2011). Secondary preventive medication persistence and adherence 1 year after stroke. *Neurology*, 77(12): 1182–1190.

Cain, LE, & Cole, SR. (2009). Inverse probability-of-censoring weights for the correction of time-varying noncompliance in the effect of randomized highly active antiretroviral therapy on incident AIDS or death. *Stat Med*, 28(12): 1725–1738.

Cavallari, LH, Aston, JL, Momary, KM, Shapiro, NL, Patel, SR, & Nutescu, EA. (2009). Predictors of unstable anticoagulation in African Americans. *J Thromb Thrombolysis*, 27(4): 430–437.

Cavallari, LH, Butler, C, Langaee, TY, et al. (2011). Association of apolipoprotein E genotype with duration of time to achieve a stable warfarin dose in African-American patients. *Pharmacotherapy*, 31(8): 785–792.

Cavallari, LH, Langaee, TY, Momary, KM, et al. (2010). Genetic and clinical predictors of warfarin dose requirements in African Americans. *Clin Pharmacol Ther*, 87(4): 459–464.

Charlton, B, & Redberg, R. (2014). The trouble with dabigatran. *BMJ*, 349: g4681.

Cohen, D. (2014a). Concerns over data in key dabigatran trial. *BMJ*, 349: g4747.

Cohen, D. (2014b). Dabigatran: how the drug company withheld important analyses. *BMJ*, 349: g4670.

Concato, J, Peduzzi, P, Holford, TR, & Feinstein, AR. (1995). Importance of events per independent variable in proportional hazards analysis I. Background, goals, and general strategy. *J Clin Epidemiol*, 48(12): 1495–1501.

Connolly, SJ, Ezekowitz, MD, Yusuf, S, et al. (2009). Dabigatran versus warfarin in patients with atrial fibrillation. *N Engl J Med*, 361(12): 1139–1151.

Cotton, BA, McCarthy, JJ, & Holcomb, JB. (2011). Acutely injured patients on dabigatran. *N Engl J Med*, 365(21): 2039–2040.

Cove, CL, & Hylek, EM. (2013). An updated review of target-specific oral anticoagulants used in stroke prevention in atrial fibrillation, venous thromboembolic disease, and acute coronary syndromes. *J Am Heart Assoc*, 2(5): e000136.

Crews, KR, Hicks, JK, Pui, C-H, Relling, M V, & Evans, WE. (2012). Pharmacogenomics and individualized medicine: translating science into practice. *Clin Pharmacol Ther*, 92(4): 467–475.

Cutler, TW, Chuang, A, Huynh, TD, et al. (2014). A retrospective descriptive analysis of patient adherence to dabigatran at a large academic medical center. *J Manag Care Pharm*, 20(10): 1028–1034.

Dantas, GC, Thompson, B V, Manson, JA, Tracy, CS, & Upshur, REG. (2004). Patients' perspectives on taking warfarin: qualitative study in family practice. *BMC Fam Pract*, 5: 15.

Dorie, V. (2014). blme: Bayesian Linear Mixed-Effects Models.

Ediger, JP, Walker, JR, Graff, L, et al. (2007). Predictors of medication adherence in inflammatory bowel disease. *Am J Gastroenterol*, 102(7): 1417–1426.

Efron, B, & Tibshirani, RJ. (1994). *An Introduction to the Bootstrap*. CRC Press.

Fang, MC, Go, AS, Chang, Y, et al. (2010). Warfarin discontinuation after starting warfarin for atrial fibrillation. *Circ Cardiovasc Qual Outcomes*, 3(6): 624–631.

Fihn, SD, McDonell, M, Martin, D, et al. (1993). Risk factors for complications of chronic anticoagulation. A multicenter study. Warfarin Optimized Outpatient Follow-up Study Group. *Ann Intern Med*, 118(7): 511–520.

Finkelman, BS, Gage, BF, Johnson, JA, Brensinger, CM, & Kimmel, SE. (2011). Genetic warfarin dosing: tables versus algorithms. *J Am Coll Cardiol*, 57(5): 612–618.

Fitzmaurice, GM, Laird, NM, & Ware, JH. (2011). *Applied Longitudinal Analysis* (2nd ed.). Hoboken, NJ: Wiley.

Freeman, J V, Zhu, RP, Owens, DK, et al. (2011). Cost-effectiveness of dabigatran compared with warfarin for stroke prevention in atrial fibrillation. *Ann Intern Med*, 154(1): 1–11.

Gage, BF, Eby, C, Johnson, JA, et al. (2008). Use of pharmacogenetic and clinical factors to predict the therapeutic dose of warfarin. *Clin Pharmacol Ther*, 84(3): 326–331.

Gage, BF, Eby, C, Milligan, PE, Banet, GA, Duncan, JR, & McLeod, HL. (2004). Use of pharmacogenetics and clinical factors to predict the maintenance dose of warfarin. *Thromb Haemost*, 91(1): 87–94.

Gage, BF, Yan, Y, Milligan, PE, et al. (2006). Clinical classification schemes for predicting hemorrhage: results from the National Registry of Atrial Fibrillation (NRAF). *Am Hear J*, 151(3): 713–719.

Granger, CB, Alexander, JH, McMurray, JJ, et al. (2011). Apixaban versus warfarin in patients with atrial fibrillation. *N Engl J Med*, 365(11): 981–992.

Groenwold, RHH, Donders, ART, Roes, KCB, Harrell, FE, & Moons, KGM. (2012). Dealing with missing outcome data in randomized trials and observational studies. *Am J Epidemiol*, 175(3): 210–217.

Gullov, AL, Koefoed, BG, & Petersen, P. (1999). Bleeding during warfarin and aspirin therapy in patients with atrial fibrillation: the AFASAK 2 study. Atrial Fibrillation Aspirin and Anticoagulation. *Arch Intern Med*, 159(12): 1322–1328.

Hankey, GJ, & Eikelboom, JW. (2010). Antithrombotic drugs for patients with ischaemic stroke and transient ischaemic attack to prevent recurrent major vascular events. *Lancet Neurol*, 9(3): 273–284.

Hastie, T, Tibshirani, R, & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer Series in Statistics.

Heagerty, PJ, Lumley, T, & Pepe, MS. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2): 337–344.

Heagerty, PJ, & Saha-Chaudhuri, P. (2013). survivalROC: Time-dependent ROC curve estimation from censored survival data.

Hickey, GL, Grant, SW, Caiado, C, et al. (2013). Dynamic prediction modeling approaches for cardiac surgery. *Circ Cardiovasc Qual Outcomes*, 6(6): 649–658.

Higashi, MK, Veenstra, DL, Kondo, LM, et al. (2002). Association between CYP2C9 genetic variants and anticoagulation-related outcomes during warfarin therapy. *JAMA*, 287(13): 1690–1698.

Hoeting, JA, Madigan, D, Raftery, AE, & Volinsky, CT. (1999). Bayesian Model Averaging: a Tutorial. *Stat Sci*, 14(4): 382–417.

Hylek, EM, Heiman, H, Skates, SJ, Sheehan, MA, & Singer, DE. (1998). Acetaminophen and other risk factors for excessive warfarin anticoagulation. *JAMA*, 279(9): 657–662.

Hylek, EM, Skates, SJ, Sheehan, MA, & Singer, DE. (1996). An analysis of the lowest effective intensity of prophylactic anticoagulation for patients with nonrheumatic atrial fibrillation. *N Engl J Med*, 335(8): 540–546.

Jaeger, M, Jeanneret, B, & Schaeren, S. (2011). Spontaneous spinal epidural haematoma during Factor Xa inhibitor treatment (Rivaroxaban). *Eur Spine J*, 21(Suppl 4): S433–S435.

Jorgensen, AL, Al-Zubiedi, S, Zhang, JE, et al. (2009). Genetic and environmental factors determining clinical outcomes and cost of warfarin therapy: a prospective study. *Pharmacogenet Genomics*, 19(10): 800–812.

Justice, AC, Covinsky, KE, & Berlin, JA. (1999). Assessing the Generalizability of Prognostic Information. *Ann Intern Med*, 130(6): 515–524.

Kanagasabapathy, P, Chowdary, P, & Gatt, A. (2011). Alternatives to Warfarin-The Next Generation of Anticoagulants. *Cardiovasc Ther*, 29(6): e80–e88.

Kealey, C, Chen, Z, Christie, J, et al. (2007). Warfarin and cytochrome P450 2C9 genotype: possible ethnic variation in warfarin sensitivity. *Pharmacogenomics*, 8(3): 217–225.

Kerr, KF, McClelland, RL, Brown, ER, & Lumley, T. (2011). Evaluating the incremental value of new biomarkers with integrated discrimination improvement. *Am J Epidemiol*, 174(3): 364–374.

Kimmel, SE, Chen, Z, Price, M, et al. (2007). The influence of patient adherence on anticoagulation control with warfarin: results from the International Normalized Ratio Adherence and Genetics (IN-RANGE) Study. *Arch Intern Med*, 167(3): 229–235.

Kimmel, SE, Christie, J, Kealey, C, et al. (2008). Apolipoprotein E genotype and warfarin dosing among Caucasians and African Americans. *Pharmacogenomics J*, 8(1): 53–60.

Kimmel, SE, French, B, Kasner, SE, et al. (2013). A Pharmacogenetic versus a Clinical Algorithm for Warfarin Dosing. *N Engl J Med*, 369(24): 2283–2293.

Klein, TE, Altman, RB, Eriksson, N, et al. (2009). Estimation of the warfarin dose with clinical and pharmacogenetic data. *N Engl J Med*, 360(8): 753–764.

König, IR, Malley, JD, Weimar, C, Diener, H-C, & Ziegler, A. (2007). Practical experiences on the necessity of external validation. *Stat Med*, 26(30): 5499–5511.

Kukafka, R, Ancker, JS, Chan, C, et al. (2007). Redesigning electronic health record systems to support public health. *J Biomed Inform*, 40(4): 398–409.

Kulkarni, SP, Alexander, KP, Lytle, B, Heiss, G, & Peterson, ED. (2006). Long-term adherence with cardiovascular drug regimens. *Am Heart J*, 151(1): 185–191.

Lee, MTM, & Klein, TE. (2013). Pharmacogenetics of warfarin: challenges and opportunities. *J Hum Genet*, 58(6): 334–338.

Lenzini, P, Wadelius, M, Kimmel, S, et al. (2010). Integration of genetic, clinical, and INR data to refine warfarin dosing. *Clin Pharmacol Ther*, 87(5): 572–578.

Li, N, Elashoff, RM, & Li, G. (2009). Robust joint modeling of longitudinal measurements and competing risks failure time data. *Biom J*, 51(1): 19–30.

Limdi, NA, Arnett, DK, Goldstein, JA, et al. (2008). Influence of CYP2C9 and VKORC1 on warfarin dose, anticoagulation attainment and maintenance among European-Americans and African-Americans. *Pharmacogenomics*, 9(5): 511–526.

Limdi, NA, Wiener, H, Goldstein, JA, Acton, RT, & Beasley, TM. (2009). Influence of CYP2C9 and VKORC1 on warfarin response during initiation of therapy. *Blood Cells Mol Dis*, 43(1): 119–128.

Lip, GY, Frison, L, Halperin, JL, & Lane, DA. (2011). Comparative validation of a novel risk score for predicting bleeding risk in anticoagulated patients with atrial fibrillation: the HAS-

BLED (Hypertension, Abnormal Renal/Liver Function, Stroke, Bleeding History or Predisposition, Labile INR, Elderly, Drug. *J Am Coll Cardiol*, 57(2): 173–180.

Liu, M, Kapadia, AS, & Etzel, CJ. (2010). Evaluating a New Risk Marker's Predictive Contribution in Survival Models. *J Stat Theory Pract*, 4(4): 845–855.

Lu, G, DeGuzman, FR, Hollenbach, SJ, et al. (2013). A specific antidote for reversal of anticoagulation by direct and indirect inhibitors of coagulation factor Xa. *Nat Med*, 19(4): 446–451.

Ma, Q, & Lu, AYH. (2011). Pharmacogenetics, pharmacogenomics, and individualized medicine. *Pharmacol Rev*, 63(2): 437–459.

Mangiafico, RA, & Mangiafico, M. (2012). Emerging Anticoagulant Therapies for Atrial Fibrillation : New Options , New Challenges. *Curr Med Chem*, 19: 4688–4698.

McCormick, TH, Raftery, AE, Madigan, D, & Burd, RS. (2012). Dynamic logistic regression and dynamic model averaging for binary classification. *Biometrics*, 68(1): 23–30.

McCulloch, CE, & Neuhaus, JM. (2011). Prediction of random effects in linear and generalized linear models under model misspecification. *Biometrics*, 67(1): 270–279.

Meckley, LM, Wittkowsky, AK, Rieder, MJ, Rettie, AE, & Veenstra, DL. (2008). An analysis of the relative effects of VKORC1 and CYP2C9 variants on anticoagulation related outcomes in warfarin-treated patients. *Thromb Haemost*, 100(2): 229–239.

Miller, A. (2002). *Subset Selection in Regression* (2nd ed., p. 256). Chapman and Hall/CRC.

Mohapatra, R, Tran, M, Gore, JM, & Spencer, FA. (2005). A review of the oral direct thrombin inhibitor ximelagatran: not yet the end of the warfarin era. *Am Hear J*, 150(1): 19–26.

Moons, KG, Donders, RA, Stijnen, T, & Harrell, FE. (2006). Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol*, 59(10): 1092–1101.

Moore, TJ, Cohen, MR, & Mattison, DR. (2014). Dabigatran, bleeding, and the regulators. *BMJ*, 349: g4517.

Myers, JA, Rassen, JA, Gagne, JJ, et al. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *Am J Epidemiol*, 174(11): 1213–1222.

Nathisuwan, S, Dilokthornsakul, P, Chaiyakunapruk, N, Morarai, T, Yodting, T, & Piriyachananusorn, N. (2011). Assessing evidence of interaction between smoking and warfarin: a systematic review and meta-analysis. *Chest*, 139(5): 1130–1139.

Nikolaus, T, Kruse, W, Bach, M, Specht-Leible, N, Oster, P, & Schlierf, G. (1996). Elderly patients' problems with medication. An in-hospital and follow-up study. *Eur J Clin Pharmacol*, 49(4): 255–259.

O'Dell, KM, Igawa, D, & Hsin, J. (2012). New oral anticoagulants for atrial fibrillation: a review of clinical trials. *Clin Ther*, 34(4): 894–901.

Oudega, R, Hoes, AW, & Moons, KGM. (2005). The Wells rule does not adequately rule out deep venous thrombosis in primary care patients. *Ann Intern Med*, 143(2): 100–107.

Parker, CS, Chen, Z, Price, M, et al. (2007). Adherence to warfarin assessed by electronic pill caps, clinician assessment, and patient reports: results from the IN-RANGE study. *J Gen Intern Med*, 22(9): 1254–1259.

Patel, MR, Mahaffey, KW, Garg, J, et al. (2011). Rivaroxaban versus warfarin in nonvalvular atrial fibrillation. *N Engl J Med*, 365(10): 883–891.

Pauker, SG, & Kassirer, JP. (1975). Therapeutic Decision Making: A Cost-Benefit Analysis. *N Engl J Med*, 293(5): 229–234.

Pearl, J. (2011). Invited Commentary: Understanding Bias Amplification. *Am J Epidemiol*, 174(11): 1223–1227.

Peduzzi, P, Concato, J, Feinstein, AR, & Holford, TR. (1995). Importance of events per independent variable in proportional hazards regression analysis II. Accuracy and precision of regression estimates. *J Clin Epidemiol*, 48(12): 1503–1510.

Pepe, MS, Feng, Z, & Gu, JW. (2008). Comments on "Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond" by M. J. Pencina et al., Statistics in Medicine (DOI: 10.1002/sim.2929). *Stat Med*, 27(2): 173–181.

Pepe, MS, Janes, H, Longton, G, Leisenring, W, & Newcomb, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol*, 159(9): 882–890.

Pirmohamed, M, Burnside, G, Eriksson, N, et al. (2013). A randomized trial of genotype-guided dosing of warfarin. *N Engl J Med*, 369(24): 2294–2303.

Platt, AB, Localio, AR, Brensinger, CM, et al. (2008). Risk factors for nonadherence to warfarin: results from the IN-RANGE study. *Pharmacoepidemiol Drug Saf*, 17(9): 853–860.

Platt, AB, Localio, AR, Brensinger, CM, et al. (2010). Can we predict daily adherence to warfarin?: Results from the International Normalized Ratio Adherence and Genetics (IN-RANGE) Study. *Chest*, 137(4): 883–889.

R Development Core Team. (2014). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Raftery, AE, Gneiting, T, Balabdaoui, F, & Polakowski, M. (2005). Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Mon Weather Rev*, 133(5): 1155–1174.

Ridker, PM, & Cook, NR. (2013). Statins: new American guidelines for prevention of cardiovascular disease. *Lancet*, 382(9907): 1762–1765.

Rieder, MJ, Reiner, AP, Gage, BF, et al. (2005). Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose. *N Engl J Med*, 352(22): 2285–2293.

Robins, JM, & Finkelstein, DM. (2000). Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics*, 56(3): 779–788.

Rollins, BM, Silva, MA, Donovan, JL, & Kanaan, AO. (2014). Evaluation of Oral Anticoagulants for the Extended Treatment of Venous Thromboembolism Using a Mixed-Treatment Comparison, Meta-Analytic Approach. *Clin Ther*. [Epub ahead of print].

Roman, M, Skaane, P, & Hofvind, S. (2014). The cumulative risk of false-positive screening results across screening centres in the Norwegian Breast Cancer Screening Program. *Eur J Radiol*, 83(9): 1639–1644.

Ross, JS, Mulvey, GK, Stauffer, B, et al. (2008). Statistical models and patient predictors of readmission for heart failure: a systematic review. *Arch Intern Med*, 168(13): 1371–1386.

Ryu, D, Li, E, & Mallick, BK. (2011). Bayesian nonparametric regression analysis of data with random effects covariates from longitudinal measurements. *Biometrics*, 67(2): 454–466.

Schelleman, H, Brensinger, CM, Chen, J, Finkelman, BS, Rieder, MJ, & Kimmel, SE. (2010). New genetic variant that might improve warfarin dose prediction in African Americans. *Br J Clin Pharmacol*, 70(3): 393–399.

Schelleman, H, Chen, J, Chen, Z, et al. (2008). Dosing algorithms to predict warfarin maintenance dose in Caucasians and African Americans. *Clin Pharmacol Ther*, 84(3): 332–339.

Schelleman, H, Chen, Z, Kealey, C, et al. (2007). Warfarin response and vitamin K epoxide reductase complex 1 in African Americans and Caucasians. *Clin Pharmacol Ther*, 81(5): 742–747.

Schelleman, H, Limdi, NA, & Kimmel, SE. (2008). Ethnic differences in warfarin maintenance dose requirement and its relationship with genetics. *Pharmacogenomics*, 9(9): 1331–1346.

Schootman, M, Lian, M, Pruitt, SL, et al. (2014). Hospital and geographic variability in two colorectal cancer surgery outcomes: complications and mortality after complications. *Ann Surg Oncol*, 21(8): 2659–2666.

Schwarz, UI, Ritchie, MD, Bradford, Y, et al. (2008). Genetic determinants of response to warfarin during initial anticoagulation. *N Engl J Med*, 358(10): 999–1008.

Sconce, EA, Khan, TI, Wynne, HA, et al. (2005). The impact of CYP2C9 and VKORC1 genetic polymorphism and patient characteristics upon warfarin dose requirements: proposal for a new dosing regimen. *Blood*, 106(7): 2329–2333.

Shah, S V, & Gage, BF. (2011). Cost-effectiveness of dabigatran for stroke prophylaxis in atrial fibrillation. *Circulation*, 123(22): 2562–2570.

Sherman, BW, Sekili, A, Prakash, ST, & Rausch, CA. (2011). Physician-specific variation in medication adherence among diabetes patients. *Am J Manag Care*, 17(11): 729–736.

Shireman, TI, Mahnken, JD, Howard, PA, Kresowik, TF, Hou, Q, & Ellerbeck, EF. (2006). Development of a contemporary bleeding risk model for elderly warfarin recipients. *Chest*, 130(5): 1390–1396.

Siegal, DM, & Crowther, MA. (2013). Acute management of bleeding in patients on novel oral anticoagulants. *Eur Heart J*, 34(7): 489–498b.

Silber, JH, Rosenbaum, PR, Brachet, TJ, et al. (2010). The Hospital Compare mortality model and the volume-outcome relationship. *Health Serv Res*, 45(5p1): 1148–1167.

Siontis, GCM, Tzoulaki, I, Castaldi, PJ, & Ioannidis, JPA. (2014). External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol*. [Epub ahead of print].

Song, X, Sander, SD, Varker, H, & Amin, A. (2012). Patterns and predictors of use of warfarin and other common long-term medications in patients with atrial fibrillation. *Am J Cardiovasc Drugs*, 12(4): 245–253.

Steffel, J, & Braunwald, E. (2011). Novel oral anticoagulants: focus on stroke prevention and treatment of venous thrombo-embolism. *Eur Hear J*, 32(16): 1968–1976a.

Steyerberg, EW, Eijkemans, MJC, Harrell, FEJ, & Habbema, JDF. (2000). Prognostic modelling with logistic regression analysis : a comparison of selection and estimation methods in small data sets. *Stat Med*, 19(8): 1059–1079.

Stone, NJ, Robinson, JG, Lichtenstein, AH, et al. (2014). 2013 ACC/AHA guideline on the treatment of blood cholesterol to reduce atherosclerotic cardiovascular risk in adults: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *J Am Coll Cardiol*, 63(25 Pt B): 2889–2934.

Suarez-Kurtz, G, & Botton, MR. (2013). Pharmacogenomics of warfarin in populations of African descent. *Br J Clin Pharmacol*, 75(2): 334–346.

Uchino, K, & Hernandez, AV. (2012). Dabigatran association with higher risk of acute coronary events: meta-analysis of noninferiority randomized controlled trials. *Arch Intern Med*, 172(5): 397–402.

Van der Heijden, GJ, Donders, AR, Stijnen, T, & Moons, KG. (2006). Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. *J Clin Epidemiol*, 59(10): 1102–1109.

Van Houwelingen, JC, & Le Cessie, S. (1990). Predictive value of statistical models. *Stat Med*, 9(11): 1303–1325.

Vickers, AJ, Cronin, AM, Elkin, EB, & Gonen, M. (2008). Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. *BMC Med Inform Decis Mak*, 8: 53.

Vickers, AJ, & Elkin, EB. (2006). Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Mak*, 26(6): 565–574.

Voora, D, Eby, C, Linder, MW, et al. (2005). Prospective dosing of warfarin based on cytochrome P-450 2C9 genotype. *Thromb Haemost*, 93(4): 700–705.

Wadelius, M, Sorlin, K, Wallerman, O, et al. (2004). Warfarin sensitivity related to CYP2C9, CYP3A5, ABCB1 (MDR1) and other factors. *Pharmacogenomics J*, 4(1): 40–48.

White, RH, Hong, R, Venook, AP, et al. (1987). Initiation of warfarin therapy: comparison of physician dosing with computer-assisted dosing. *J Gen Intern Med*, 2(3): 141–148.

Wieloch, M, Själander, A, Frykman, V, Rosenqvist, M, Eriksson, N, & Svensson, PJ. (2011). Anticoagulation control in Sweden: reports of time in therapeutic range, major bleeding, and thrombo-embolic complications from the national quality registry AuriculA. *Eur Heart J*, 32(18): 2282–2289.

Wilmott, CJ, & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim Res*, 30: 79–82.

Witt, DM, Delate, T, Clark, NP, et al. (2009). Outcomes and predictors of very stable INR control during chronic anticoagulation therapy. *Blood*, 114(5): 952–956.

Wysowski, DK, Nourjah, P, & Swartz, L. (2007). Bleeding complications with warfarin use: a prevalent adverse effect resulting in regulatory action. *Arch Intern Med*, 167(13): 1414–1419.

Yap, C-H, Reid, C, Yii, M, et al. (2006). Validation of the EuroSCORE model in Australia. *Eur J Cardiothorac Surg*, 29(4): 441–446.